AD_____

GRANT NUMBER DAMD17-94-J-4174

TITLE: Utilizing Serial Measures of Breast Cancer Risk Factors

PRINCIPAL INVESTIGATOR: Mimi Y. Kim, Sc.D.

CONTRACTING ORGANIZATION: New York University Medical Center
New York, New York 10010-2598

REPORT DATE: January 1999

TYPE OF REPORT: Final

PREPARED FOR: Commander
U.S. Army Medical Research and Materiel Command
Fort Detrick, Frederick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for public release;
distribution unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

DTIC QUALITY INSPECTED 4

19991026 146

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	January 1999	Final (31 Dec 94 - 30 Dec 98)

4. TITLE AND SUBTITLE
Utilizing Serial Measures of Breast Cancer Risk Factors

5. FUNDING NUMBERS
DAMD17-94-J-4174

6. AUTHOR(S)
Mimi Y. Kim, Sc.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
New York University Medical Center
New York, New York 10010-2598

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Commander
U.S. Army Medical Research and Materiel Command
Fort Detrick, Frederick, MD 21702-5012

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200*

The overall goals of the project were to develop theoretical statistical methods for utilizing repeated measurements of serum levels of risk factors in epidemiologic studies of breast cancer. During the four year study period, the following research was accomplished: (1) a technique was developed for correcting for measurement error when subjects have a variable number of repeated exposure measurements; (2) the technique was applied to studies of endogenous estrogens and androgens and risk of post-menopausal breast cancer; (3) a method was developed for adjusting for the systematic variability of hormone levels over the menstrual cycle in studies of pre-menopausal breast cancer; (4) guidelines were devised for choosing the number of repeated measurements and optimal time interval between repeats in half-life studies of environmental contaminants which have been linked with an increased risk of breast cancer; (5) methodology was developed for analyzing correlated panel data which can arise when serially monitoring toxicity in clinical trials of breast cancer treatments; and (6) a conceptual approach based on equivalence testing was proposed for assessing the validity of a dietary assessment method for use in epidemiologic studies of breast cancer and other diseases.

14. SUBJECT TERMS
Biostatistics, Hormones, Repeated Measures, Measurement Error, Time-Dependent Covariates, Breast Cancer

15. NUMBER OF PAGES
119

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

____ Where copyrighted material is quoted, permission has been obtained to use such material.

____ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

____ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

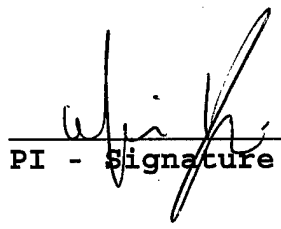
____ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

____ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

____ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

____ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

____ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.



PI - Signature 2/25/91 Date

Overview

This **final report** describes research accomplished during the period December 31, 1994 - December 30, 1998 on the project, **"Utilizing serial measures of breast cancer risk factors".** The overall goals of the project were to develop and apply theoretical statistical methods for utilizing repeated measurements of serum levels of endogenous hormones and other biologic measurements in epidemiologic studies of breast cancer. During the four year study period, the following research was accomplished: (1) a technique was developed for correcting for measurement error when subjects have a variable number of repeated exposure measurements; (2) the technique was applied to nested case-control studies of endogenous estrogens and androgens and risk of post-menopausal breast cancer; (3) a method was developed for adjusting for the systematic variability of hormone levels over the menstrual cycle in order to evaluate the association between estrogen levels and risk of pre-menopausal breast cancer; (4) guidelines were devised for choosing the number of repeated measurements and optimal time interval between repeats in half-life studies of environmental contaminants which have been linked with an increased risk of breast cancer; (5) methodology was developed for analyzing correlated panel data which can arise when serially monitoring toxicity in clinical trials of breast cancer treatments; and (6) a conceptual approach based on equivalence testing was proposed for assessing the validity of a dietary assessment method for use in epidemiologic studies of breast cancer and other diseases.

This report is organized into six chapters detailing the background, methods, results and conclusions of each of the six research areas described above. Four manuscripts based on this project have been published (see Bibliography), a fifth manuscript has been submitted for publication, and a sixth is in preparation. Not only were the main goals outlined in the Statement of Work achieved, but additional research beyond the scope of the original proposal was also successfully executed. This career development award has yielded important methodologic contributions to the analysis of breast cancer studies involving repeated measurements of potential risk factors and has enabled the principal investigator to establish a solid foundation for pursuing further research in this area.

Table of Contents

1. Correcting for Measurement Error in the Analysis of Case-Control Data with Repeated Measurements of Exposure

1.1 Introduction....8
1.1 Methods....10
1.2 Example....15
1.3 Conclusions....18
1.4 References....22
1.5 Appendix....27

2. Relation of Serum Testosterone and Dehydroepiandrosterone Sulfate with Risk of Breast Cancer in Postmenopausal Women

2.1 Introduction....30
2.2 Methods....32
2.3 Results....36
2.4 Conclusions....38
2.5 References....43

3. Sample Size and Study Design Considerations for Half-Life Studies

3.1 Introduction....55
3.2 Methods....56
3.3 Example....62
3.4 Conclusions....65
3.5 References....70
3.6 Appendix I....72
3.7 Appendix II....74

4. The Analysis of Correlated Panel Data Using a Continuous Time Markov Model

4.1 Introduction....77
4.2 Methods....78
4.3 Example....82
4.4 Conclusions....86
4.5 References....87

5. The Application of Equivalence Tests in Validation Studies of Dietary Assessment Methods

5.1 Introduction....92
5.2 Methods....95
5.3 Conclusions....102
5.4 References....104

6. Adjusting Hormone Levels for Day of Menstrual Cycle in Studies of Hormones and Pre-Menopausal Risk of Breast Cancer

6.1 Introduction.....107
6.2 Methods.....108
6.3 Results.....111
6.4 Conclusions.....114
6.5 References.....115

7. Bibliography.....119

Chapter I

Correcting for Measurement Error in the Analysis of Case-Control Data with Repeated Measurements of Exposure

1 Introduction

In most case-control studies, the risk factors of interest are measured with error. For biologic variables, such as blood pressure, nutrient, and hormone levels, measurement error can arise from limitations in the measurement technique or laboratory assay. In addition, because the exposure of interest is usually a subject's underlying long-term average value rather than the level at any single point in time, intrinsic fluctuations in the variable over time can also contribute to measurement error.

When the error is random and non-differential with respect to case-control status, it is well known that estimates of relative risk based on the mis-measured exposure will be attenuated. In order to minimize the effects of measurement error, many investigators advocate collecting repeated measurements of the exposure on all subjects and using the individual's average value (1). However, as noted by Rosner et al (2), even when the mean of several replicates is substituted for a single measurement, attenuation of relative risks may still occur, especially when the degree of measurement error is large and the average is based on only a few repeats.

Methods for correcting estimates of relative risk for measurement error have been proposed in a number of epidemiologic and statistical papers (3,4). The most common method involves correcting the "naive" relative risk estimate based on the observed exposure by the expected amount of bias. In the case of logistic regression, the regression parameter will be attenuated by the factor, R, which is equal to the reliability coefficient of the mis-measured exposure (1, 2). Therefore, one can multiply the biased estimate of the regression coefficient by the inverse of the reliability coefficient to obtain the corrected estimate. This method, however, is dependent on the assumption that the reliability of the exposure measurement is the same for all subjects. When the average of several replicates is used as the measure of exposure, this condition will be met only if all subjects have an equal number of repeated measurements, given the degree of measurement error associated with a single measurement is the same for all subjects.

In studies in which the exposure is measured on repeated occasions, however, subjects often have a variable number of measurements because of missing data. For example, the data that are utilized to illustrate the methods in this paper are derived from a nested case-control study of serum hormonal levels and breast cancer from the NYU Women's Health Study (5). The study cohort consists of 14,275 women who donated multiple blood samples over time and have been followed since enrollment for the development of breast cancer. Most women have donated one or two samples; however, many have also donated three or more. Because subjects with a larger number of multiple blood samples have a more reliable estimate of their true underlying serum hormonal levels than those with fewer measurements, the reliability of the measured exposure will not be constant across individuals. Consequently, the usual procedure for correcting for measurement error cannot be applied.

Liu and Liang (6) proposed an estimating equation approach for obtaining consistent estimates of logistic regression parameters when all subjects have the same number of repeated imprecise exposure measurements, which in principle could be extended to the more complicated situation when the number of replicates is variable between subjects. In this paper, we discuss an alternative method for correcting for measurement error in the analysis of matched case-control data when subjects have a variable number of repeated exposure measurements and the individual's average is used as the measure of exposure. The technique, which assumes that both the true exposure and the measurement error are normally distributed, involves multiplying each subject's observed average by the reliability of the average prior to fitting the logistic regression model. The resulting logistic regression coefficient based on the transformed average is corrected for measurement error. A bootstrap algorithm for obtaining confidence intervals for the regression parameter which takes into account the variability due to estimation of the reliability coefficient is also proposed.

2 Methods

Measurement Error Model and Correction of Logistic Regression Parameter

The methods described below are based on the measurement error model of Armstrong et al (7) for matched case-control studies. We assume that in each matching stratum, a case is matched to a variable number of controls. However, the techniques are generalizable to the unmatched design by assuming that there is only one matching stratum.

Let x_{ijk} denote the unobserved true value of the exposure variable for the k^{th} subject with case/control status j (0 = control, 1 = case), in stratum i ($i = 1, ..., M$). Assume that x_{ijk} is normally distributed with mean, $\mu_i + j\delta$, and variance σ_s^2. In addition, let z_{ijkl} denote the l^{th} observed value of x_{ijk}, measured with error, for $l = 1, ..., n_{ijk}$. We assume the following classical errors-in-variables model:

$$z_{ijkl} = x_{ijk} + e_{ijkl},$$

where the error term, e_{ijkl}, is independent of x_{ijk} and $e_{ijkl'}$, for $l \neq l'$, and normally distributed with mean 0 and variance, σ_e^2. It follows that the observed z_{ijkl} in stratum i are normally distributed with means $\mu_i + \delta$ and μ_i for cases and controls, respectively, and common variance, $\sigma_s^2 + \sigma_e^2$. The variance component, σ_s^2, can be interpreted as the variance of the true exposure, after stratifying by matching stratum and case/control status, and σ_e^2 as the variance due to measurement error.

With these assumptions and the application of Bayes' Rule, Armstrong et al (7) showed that the probability that a study subject is a case, conditional on $\bar{z}_n$, the observed average based on n measurements, and membership in stratum i, is a logistic function:

$$\Pr(D = 1|\bar{z}_n; i) = \frac{\exp(\alpha_i + \beta R_n \bar{z}_n)}{1 + \exp(\alpha_i + \beta R_n \bar{z}_n)}, \tag{1}$$

where

$$R_n = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2/n} \tag{2}$$

is the reliability of $\bar{z}_n$ as a measure of x. When no measurement error is present, $\bar{z}_n = x$, the reliability coefficient is equal to 1, and (1) reduces to:

$$\Pr(D = 1|x; i) = \frac{\exp(\alpha_i + \beta x)}{1 + \exp(\alpha_i + \beta x)}.$$

Thus, an estimate of the logistic regression coefficient based on $\bar{z}_n$ will estimate the "naive" coefficient, $\beta^* = \beta R_n$, rather than the true β. Because the reliability coefficient is between 0 and 1, the "naive" β^* will be attenuated relative to β. We can see from (2), however, that as the number of repeated measurements increases, the reliability coefficient approaches 1, and the corresponding attenuation in β will diminish.

When all subjects have the same number of n repeated measurements, an estimate of the true regression coefficient can be obtained by fitting the logistic model using $\bar{z}_n$ for each subject's exposure measurement, and multiplying the resulting coefficient estimate, $\hat{\beta}^*$, by $1/R_n$. If subjects have a variable number of measurements, however, this approach cannot be applied, since the reliability of the exposure variable is no longer constant for all subjects, but depends on the number of available repeated measurements.

For the case where the reliability of the exposure differs across subjects, a corrected estimate of the regression coefficient may be obtained by multiplying each subject's average exposure measurement by the reliability of the average, prior to model fitting. That is, if the k^{th} subject in stratum i has the observed average $\bar{z}_{ijk.}$, based on n_{ijk} approximate measurements of x_{ijk}, then replacing the unknown x_{ijk} in the conditional logistic model with the transformed average, $R_{n_{ijk}}\bar{z}_{ijk.}$, where $R_{n_{ijk}}$ is calculated from (2), will yield an estimate of the true β. Since the reliability increases with the number of measurements, this transformation results in greater "shrinkage" of averages based on a small number of repeats, and less shrinkage of more informative averages based on many repeats.

When all subjects within the same matched set have the same number of repeats, this method is equivalent to the two-stage approach proposed by Thomas et al (4) and Whittemore (8) for error correction in linear models, in which $E(x_{ijk}|\bar{z}_{ijk.})$ is computed and then

used as the exposure in the usual regression model. Given the model assumptions described above, $E(x_{ijk}|\bar{z}_{ijk.}) = R_{n_{ijk}}\bar{z}_{ijk.} + (1 - R_{n_{ijk}})E(x_{ijk})$. If n_{ijk} is constant for all subjects in stratum i, the $(1 - R_{n_{ijk}})E(x_{ijk})$ term is absorbed in the intercept term and does not affect the estimate of the slope parameter in the logistic regression model. Thus, utilizing $E(x_{ijk}|\bar{z}_{ijk.})$ or $R_{n_{ijk}}\bar{z}_{ijk.}$ will yield equivalent estimates of the true regression parameter. Furthermore, when all subjects in the study have the same number of n repeats, this technique will result in a corrected estimate of the logistic regression coefficient that is identical to the one obtained by correcting the naive estimate by $1/R_n$.

Although fitting the logistic model to the transformed covariate will result in an unbiased estimate of β, the corresponding variance of $\hat{\beta}$ will be underestimated unless the variance components in the reliability coefficient are known. Usually, however, the variance components are estimated from a separate reliability substudy or from subjects in the main study. In our setting, repeated measurements are assumed to be available on all or a subset of the main study participants. Thus, we can estimate the variance components, σ_s^2 and σ_e^2, from the main study data by fitting the following mixed effects analysis of variance model to data on all cases and controls:

$$z_{ijkl} = \mu_i + \delta_j + \gamma_{ijk} + \epsilon_{ijkl}, \tag{3}$$

where z_{ijkl} is defined as before, μ_i is the effect for stratum i, δ_j is the effect due to case/control status, γ_{ijk} is a normally distributed random subject effect with mean 0 and variance σ_s^2, and ϵ_{ijkl} is the residual error which is normally distributed with mean 0 and variance σ_e^2.

The variance components, σ_s^2 and σ_e^2, can be estimated using one of several methods, including traditional analysis of variance (ANOVA), maximum likelihood, or restricted maximum likelihood methods. The ANOVA method, available in the SAS procedure, PROC GLM, was used in our example because it is computationally simpler than the others, which is an important consideration when implementing the bootstrap procedure described in the next section for generating confidence intervals. However, this technique can lead to negative

variance estimates. The ML or REML estimators, which are available from PROC MIXED in SAS, do not have this limitation. For further details about the different estimation techniques, see Searle, Casella and McCulloch (9).

The steps involved in obtaining an estimate of the logistic regression coefficient corrected for measurement error can be summarized as follows:

1. Estimate the variance components, σ_e^2 and σ_s^2, by fitting the mixed effects analysis of variance model in (3) to the study data.

2. Multiply each subject's average exposure by $\hat{R}_{n_{ijk}} = \hat{\sigma}_s^2/(\hat{\sigma}_s^2 + \hat{\sigma}_e^2/n_{ijk})$.

3. Estimate the true logistic regression coefficient, β, by fitting a conditional logistic regression model to the transformed averages.

Because the technique is based on assuming that the true exposure and measurement error are normally distributed, suitable data transformations should be applied when the distributions deviate from normality. Note, however, that a data transformation such as the log-transform will result in a model in which the log odds of disease is a linear function of the exposure measured on the log, rather than the original scale.

Bootstrap Method for Obtaining Confidence Intervals

The width of the usual 95% confidence interval for the true β based on the transformed covariate will be too narrow because the interval does not account for the extra variability associated with estimation of the variance components in R_n. Rosner et al. (2) have derived the asymptotic variance and corresponding confidence intervals of the corrected logistic regression parameter which includes the uncertainty of the variance estimates for use in cohort studies under a rare disease assumption. Their method, however is applicable only when all subjects in the main study have the same number of repeats. For the situation when

subjects in a matched case-control study have a variable number of replicates, we propose the following bootstrap algorithm for obtaining confidence intervals for the true β:

1. Assuming there are M matched sets in the main study, generate a bootstrap sample using the matching stratum as the sampling unit, and sampling M matched sets with replacement from the main study data. For each matched set that is selected, the sample contains all the subjects within the matched set, along with each subject's case/control status and repeated measurements.

2. Using the bootstrap sample, estimate σ_s^2, σ_e^2, and the true β by following the 3-step approach outlined in the previous section.

3. Repeat (1) and (2) 1,000 times, which is the approximate minimum number of bootstraps necessary to compute bias-corrected confidence limits (10).

In constructing a bootstrap sample from the main study data, sampling occurs at the level of the matching stratum since the matching between the cases and the controls needs to be preserved. If the number of controls matched to each case is variable across strata, one can sample the strata according to the number of subjects in each matched set, in order to keep the total sample size constant for each bootstrap iteration. For example, one samples with replacement M_2 matched sets from the M_2 sets in the main study with 2 controls per case, M_3 sets from the strata with 3 controls per case, etc...

The simple $(1-\alpha)\%$ confidence interval can be constructed using the $\alpha/2$ and $(1-\alpha/2)$ percentiles of the bootstrap distribution. Bias-corrected confidence intervals should be used when the bootstrap distribution of β is asymmetric and when the sample size is small (10). We report only the bias-corrected confidence intervals in this paper.

Thus far, our focus has been on correcting for measurement error in a single exposure variable, in the absence of confounders. However, the methods can also be generalized to the

multi-covariate situation, where the confounders, in addition to the primary exposure variable, may be measured with error. A brief outline of the methods is given in the Appendix. Additional details on the measurement error model and estimation of variance components are also described in Armstrong et al (7).

3 Example

The primary aim of the NYU Women's Health Study is to determine whether serum levels of endogenous hormones, such as estradiol, are associated with risk of breast cancer. Between March 1985 and June 1991, a cohort of 14,275 healthy women aged 34-65 years were enrolled at the Guttman Breast Diagnostic Institute, New York. At the time of enrollment and at annual screening visits thereafter, women were asked to donate blood and complete a self-administered questionnaire. Serum samples were frozen and stored for future biological assays. Subsequent cases of breast cancer were identified primarily through active follow-up and confirmed by reviewing medical and pathological records. In this example, only the women who were post-menopausal at enrollment (49%) were included.

In order to limit the costs associated with measuring hormone levels in the cohort, a nested case-control study design was used. For each incident case of breast cancer, individually matched controls were selected at random from the risk set consisting of all cohort members alive and free of breast cancer at the time of diagnosis of the case, and who matched the case on menopausal status at entry, age at entry, and number and approximate dates of blood donations up to the date of diagnosis in the case. For additional details of the study design, see Toniolo et al (5).

The goal of this example is to evaluate the effect of random measurement error on the associations between total, % free, and % bound to sex hormone binding globulin (SHBG-bound) estradiol levels and risk of breast cancer, when the average of all the available repeated measurements for a subject is used as her exposure. The associations between the

baseline measurements of the total, % free, and % SHBG-bound estradiol levels and risk of breast cancer among post-menopausal women, unadjusted for measurement error, were evaluated by Toniolo et al (11). Total and % free estradiol were found to be positively associated with risk of breast cancer, whereas % SHGB-bound estradiol had a strong protective effect.

One of the assumptions of the measurement error model is that the true and observed exposure variables are normally distributed. The distribution of total estradiol levels was skewed, so the logarithm of the values were used. Based on data from both post-menopausal cases and controls, we estimated the reliability coefficients for total, % free and % SHBG-bound estradiol, adjusted for matching stratum and case/control status, as: .48, .68, and .92, respectively (Table 1). (These estimates were somewhat lower than those published by Toniolo et al (12): .51, .77, and .94 for total, % free and % SHBG-bound estradiol, respectively, which were based on data from only the post-menopausal controls in the NYU Women's Health Study.) The estimates of the reliability coefficients indicate that the degree of measurement error in total and % free estradiol may be sufficiently large to attenuate observed relationships with risk of breast cancer.

The main case-control study sample consisted of 381 subjects stratified into 130 matched sets. Ten matched sets had 1 control per case, 119 sets had 2 controls per case, and one set had 3 controls per case. Of the 381 subjects in the main study, 157 (41%) had 2 or more repeated measurements: 98 subjects had 2 replicates, 53 had 3 replicates, and 6 subjects had 4.

We investigated the effects of measurement error on the observed associations between each exposure variable and risk of breast cancer by comparing the estimated logistic regression parameters based on the first measurement of the exposure for each subject, the average of the replicate measures, and the transformed (corrected) average value. Corresponding odds ratios were calculated from the regression estimates by comparing women in the 90^{th} versus 10^{th} percentiles of the observed distributions (i.e., 63.0 vs 14.5 for total estradiol, 1.7 vs. 1.04 for % free, and 57.6 vs 27.3 for % SHBG-bound estradiol).

The bootstrap confidence intervals were generated using the SAS macro facility to create the bootstrap sample, in conjunction with PROC PHREG, which fits conditional logistic regression models. All analyses were run on a DEC 3000/700 AXP computer workstation.

The results are provided in Table 2. For total estradiol and % free estradiol, the uncorrected analyses show that using the observed average of the repeated measurements results in a minor increase in the regression coefficient estimates compared with using only the baseline measurement. On the other hand, the estimated regression coefficients corrected for measurement error using the transformed averages are substantially larger than the estimates based on the observed averages for both variables: increases are 74% and 40% for total and %free estradiol, respectively.

The effect of measurement error on the estimated odds ratios is especially striking. When comparing women in the 90th percentile versus the 10th percentile of the observed total estradiol distribution, the corrected odds ratio was estimated to be 7.16, compared with uncorrected odds ratios of 2.64 and 3.10 using the baseline and untransformed average, respectively. Similarly, the corrected odds ratio for % free estradiol was 4.95, compared with 3.07 for the baseline measurement and 3.13 for the average value.

This illustrates how using the observed average of replicate measurements of exposure for each subject may not be sufficient to offset the effects of measurement error when the degree of error is large and when subjects have only a few replicates, and that additional error correction procedures may be necessary. In the case of total estradiol, one would need to take the average of 10 replicate measurements to improve the reliability to .90, based on the estimated variance components in Table 1. For % free estradiol, one would need 5 measurements. Thus, it is not surprising that using the average value in our example did not appreciably deattenuate the corresponding regression coefficient, since only 41% of the study subjects had replicate measurements, and among these, most had only 2 or 3 measurements. Using the average resulted in a 17% increase in the regression coefficient for total estradiol, relative to using the first measurement. In comparison, if all subjects had 2 replicates, the

expected increase in the regression coefficient would be $(R_2 - R_1)/R_1 = (.65 - .48)/.48 = 37\%$ over the estimate based on 1 measurement. On the other hand, because % SHGB-bound estradiol levels are highly reproducible, the logistic regression estimates and corresponding odds ratios using the corrected average were not very different from the uncorrected analyses.

Since 119 (92%) of the 130 matched sets had 2 controls matched per case, implementation of a more complex stratified bootstrap sampling scheme, which would keep the total number of subjects constant for each iteration, was not warranted. As one would expect, the bias-corrected bootstrap confidence intervals based on the transformed average, as shown in Table 2, are shifted further away from 0 and are wider than the uncorrected confidence intervals for all variables, since the bootstrap method accounts for the variation due to estimation of the variance components in the reliability coefficient. When the variation in the estimates of variance components estimates was ignored, the simple 95% confidence intervals based on the corrected average were estimated to be: (.54, 2.13); (1.02, 3.82), and (-.074, -.025) for total, % free, and % SHBG-bound estradiol, respectively. Thus, ignoring the extra source of variation from $\hat{R}_{n_{ijk}}$ underestimated the width of the confidence interval by as much as 17% (for total estradiol) in our data set.

4 Conclusions

In most reliability studies, the within-subject or error variance of the exposure is estimated from an external population or from a random subset of the main study population from whom repeated measurements are obtained, and one must assume that the resulting estimate is generalizable to the main study population. In our example, the within-subject variances were estimated from the subjects in the main study with at least two repeated hormone measurements. Women with repeated measurements, however, may be different from those with only one measurement. Because blood samples in the NYU Women's Health Study were obtained at annual breast cancer screening visits, women with a family history of breast

cancer, for example, or those who are more health conscious, may have been more likely to return for subsequent visits. It is unlikely, though, that this would result in a systematic difference in the within-subject variability of the hormone levels between the subset with repeats and those who had only one measurement. Thus, we can assume generalizability of the estimated within-subject variance to all subjects in our main study.

A second assumption of our error-correction method is that a subject's measurements are distributed randomly around the unobserved true value, and that levels of the exposure are not changing systematically over time. This assumption may not be true if hormone levels decrease with age. In addition, for breast cancer cases, hormone levels could be influenced by the development of disease so that measurements obtained closer to the date of diagnosis may exhibit a systematic time trend. Among subjects in the NYU Women's Health Study, however, a trend in estradiol levels over time was not observed in preliminary analyses using linear regression techniques (results not shown).

We have also assumed that the variance components, σ_s^2 and σ_e^2, are homogeneous across strata and case/control status. The within-subject variance for total estradiol was estimated as .16 and .18 for cases and controls, respectively, indicating that the error variances are similar for the two groups. Because only one case was included in each stratum, we could not evaluate whether σ_s^2 was constant for cases and controls. Furthermore, assessing whether σ_s^2 was homogeneous across strata was not possible, given that most strata had only 2 controls.

The error-correction methods in this paper are applicable to studies in which a variable number of repeated measurements of exposure are obtained on subjects, and the average of each subject's measurements is used as the exposure variable. In principle, a corrected estimate of the logistic regression coefficient could also be obtained by utilizing only the first measurement of exposure for each subject, and correcting the resulting estimate by the reliability of a single measure. Although this method is much simpler than using all the available repeated measurements and applying the techniques proposed in this paper, the estimate based on a single measurement will not be as efficient. For example, the 95% bias-

corrected bootstrap confidence interval for the true β using only the first measurements of total estradiol was (.52, 2.73), which is wider by 19% than the corresponding interval based on the transformed averages.

Haukka (13) proposed a similar bootstrap method for correcting for measurement error in generalized linear models for the situation when the "gold standard" is known for the exposure measurement and validation, as opposed to reproducibility, data are available. When compared with the correction method for logistic regression proposed by Rosner et al (14) which also takes into account the variability in $\hat{R}$, the bootstrap method was found to yield wider confidence intervals for peaked and skewed measurement error distributions. As discussed by Haukka (13), this difference may result because the bootstrap method takes better account of the measurement error variance, whereas the Rosner et al. (14) method is based on a first-order Taylor series approximation, which may not adequately correct confidence intervals when the error variance is large.

We have shown that in situations when the magnitude of measurement error is large and subjects have only a few repeats, using the average of the available replicate measurements for each subject may not be sufficient to adjust for the measurement error. The methods proposed in this paper can be applied to provide additional correction procedures in the analysis of case-control data where subjects have a variable number of repeated measures of the exposure. The advantage of our algorithm is that it is conceptually straightforward and relatively easy to implement, especially with the amount of computing power that is now readily available to most investigators.

Acknowledgements. The authors are grateful to Alice Whittemore and Stuart Lipsitz for helpful comments. This work was supported by grant DAMD17-94-J-4174 from the

Department of Defense Breast Cancer Research Program and grants CA-16087 and CA-34588 from the National Institutes of Health.

References

Armstrong, B.G. The effects of measurement errors on relative risk regressions. Am J Epidemiol 1990; 132:1176-1184.

Armstrong, B., Whittemore, A., and Howe, G. Analysis of case-control data with covariate measurement error: application to diet and colon cancer. Statistics in Medicine 1989; 8:1151-1163.

De Klerk, N.H., English, D.R., and Armstrong, B.K. A review of the effects of random measurement error on relative risk estimates in epidemiological studies. International Journal of Epidemiology 1989; 18:705-712.

Efron, B. and Tibshirani, R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical Science 1986; 1:54-77.

Haukka, J.K. Correction for covariate measurement error in generalized linear models–a bootstrap approach. Biometrics 1995; 51:1127-1132.

Kim, M.Y., Pasternack, B.P., Carroll, R.J., Koenig, K.L., and Toniolo, P.G. Estimating the reliability of an exposure variable in the presence of confounders. Statistics in Medicine 1995; 14:1437-1446.

Liu, X. and Liang, K.Y. Efficacy of repeated measures in regression models with measurement error. Biometrics 1992; 48:645-654.

Prentice, R.L. Covariate measurement error and parameter estimation in a failure time

regression model. Biometrika, 1982; 69:331-342.

Rosner, B., Spiegelman, D., Willett, W.C. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. Am J Epidemiol 1990; 132:734-745.

Rosner, B., Spiegelman, D., Willett, W.C. Correction of logisitic regression relative risk estimates and confidence intervals for random within-person measurement error. Am J Epidemiol 1992; 136:1400-1413.

Thomas, D., Stram, D., and Dwyer, J. Exposure measurement error: influence on exposure-disease relationships and methods of correction. Annual Review of Public Health 1993; 14:69-93.

Toniolo, P., Pasternack, B., Shore, R., Sonnenschein, E., Koenig, K., Rosenberg, C., Strax, P., Strax, S. Endogenous hormones and breast cancer: A prospective cohort study. Breast Cancer Res Treat 1991; 18:S23-S26.

Toniolo, P., Koenig, K., Pasternack, B., Banerjee, S., Rosenberg, C., Shore, R., Strax, P., and Levitz, M. Reliability of measurements of total, protein bound, and unbound estrdiol in serum. Cancer Epidemiology, Biomarkers and Prevention 1994; 3:47-70.

Toniolo, P., Levitz, M., Zeleniuch-Jacquotte, A., Banerjee, S., Koenig, K., Shore, R., Strax, P., Pasternack, B. A prospective study of endogenous estrogens and breast cancer in post-menopausal women. Journal of the National Cancer Institute 1995; 87:190-197.

Whittemore, A.S. Errors-in-variables regression using Stein estimates. The American Statis-

tician 1989; 43:226-228.

Table 1: Reproducibility of Total, %Free, and %SHBG-Bound Estradiol, Adjusted for Case/Control Status and Matching Stratum

Hormone	Within-Subject Variance	Between-Subject Variance	Reliability Coefficient
Estradiol	0.17	0.16	0.48
% Free Estradiol	0.017	0.036	0.68
% SHBG-Bound Estradiol	9.38	104.45	0.92

Table 2: Corrected and Uncorrected Logistic Regression Parameter Estimates, Confidence Intervals, and Odds Ratios for the Associations of Total, % Free, and % SHBG-bound Estradiol Level and Risk of Breast Cancer

Exposure Variable	Regression Coefficient	95%C.I.‡	Odds Ratio*
Total Estradiol†			
First measurement	0.66	(0.24 - 1.09)	2.64
Average	0.77	(0.32 - 1.22)	3.10
Transformed average	1.34	(0.61 - 2.47)	7.16
% Free Estradiol			
First measurement	1.70	(0.69 - 2.71)	3.07
Average	1.73	(0.70 - 2.77)	3.13
Transformed average	2.42	(1.06 - 4.00)	4.95
% SHBG-Bound Estradiol			
First measurement	-0.046	(-0.068 - -0.024)	0.25
Average	-0.045	(-0.067 - -0.023)	0.26
Transformed average	-0.048	(-0.074 - -0.025)	0.24

* Comparing women at 90^{th} vs. 10^{th} percentile of observed distribution

† Total estradiol measurements were log-transformed

‡ 95% C.I. using transformed average based on bias-corrected bootstrap estimate

Appendix

In order to generalize the techniques to the multivariate situation, assume that $\mathbf{x}_{ij}$ denotes a $(p \times 1)$ vector of true covariates for the j^{th} subject in stratum i, and that it follows a multivariate normal distribution with mean vector $\mu_i + \Delta$ for the cases and μ_i for the controls, and covariance matrix Σ. In addition, let

$$\mathbf{z}_{ijk} = \mathbf{x}_{ij} + \mathbf{e}_{ijk}$$

denote the k^{th} observed measurement of $\mathbf{x}_{ij}$, for $k = 1, ...n_{ij}$, where the $\mathbf{e}_{ijk}$ are independent and identically distributed according to a multivariate normal distribution with covariance matrix, Ω.

Under these assumptions, Armstrong et al (7) showed that the probability a subject is a case, conditional on the mean of n repeated observations of the covariate vectors, is equal to the following logistic function:

$$Pr(D = 1|\bar{\mathbf{z}}_n, i) = \frac{\exp(\alpha_i + \bar{\mathbf{z}}_n \Lambda_n \beta)}{1 + \exp(\alpha_i + \mathbf{z}_n \Lambda_n \beta)},$$

where $\bar{\mathbf{z}}_n = (\sum_{k=1}^{n} \mathbf{z}_k)/n$, $\Lambda_n = (\Sigma + n^{-1}\Omega)^{-1}\Sigma$, and β is the $(p \times 1)$ vector of logistic regression parameters.

When subjects have a variable number of replicate measures of the exposure variables, it follows that as in the single covariate case, one can transform the observed mean covariate vector for each subject by multiplying the vector by an estimate of the matrix, $\Lambda_{n_{ij}}$, and then fitting the usual logistic regression model to the transformed covariates to obtain the corrected logistic regression coefficients for all covariates. A bootstrap algorithm analogous to that for the single covariate case could be used to obtain corrected confidence intervals which take into account the variation due to estimation of $\Lambda_{n_{ij}}$, but the method could become very computationally intensive with a large number of confounders, since more complicated multivariate MANOVA models would be needed to estimate Σ and Ω. For the special

case when the confounders are measured without error, however, estimation of the variance components is greatly simplified (see Kim et al (15)), and the bootstrap method can be more easily applied.

Chapter II

Relation of Serum Testosterone and Dehydroepiandrosterone Sulfate with Risk of Breast Cancer in Postmenopausal Women

1 Introduction

A possible role of androgens in the development of breast cancer in postmenopausal women was first suggested by Grattarola et al. (1974). Mechanisms by which androgens may increase breast cancer risk were reviewed by Secreto et al. (1991, 1994) and Bernstein and Ross (1993). Androgens may act directly, by stimulating breast cell proliferation through binding to androgen receptors or by stimulating the synthesis of growth factors inside the breast epithelium. Androgens may also act indirectly through their conversion to estrogens, which are known to stimulate breast cell proliferation (Henderson et al., 1993): aromatization of androstenedione and testosterone in peripheral tissues is the main source of estrogens in postmenopausal women. In addition, it is well established that testosterone binds to sex-hormone binding globulin (SHBG) with greater affinity than estradiol. Testosterone may thus indirectly increase the risk of breast cancer by decreasing the fraction of estradiol bound to SHBG and thereby increasing the nonbound fraction, which is thought to be the fraction available to breast cells (Siiteri et al., 1981). Finally, it has been suggested that testosterone inhibits hepatic secretion of SHBG (Lonning et al., 1995), which could also result in a decreased fraction of estradiol bound to SHBG.

Several case-control studies have reported on the association of plasma or serum levels of testosterone with risk of breast cancer in postmenopausal women. Most (McFadyen et al., 1976; Adami et al., 1979; Secreto et al. 1983; Hill et al.; 1985; Secreto et al., 1991; Bruning et al., unpublished data) although not all (Malarkey et al., 1977) observed higher levels of testosterone in cases than in controls. Among the three prospective studies which examined the relation of serum levels of testosterone with risk of postmenopausal breast cancer, one found a significant positive association (Berrino et al., 1996), one found a non-significant positive association (Wysowski et al., 1987) and the third reported no association (Garland et al., 1992).

Dehydroepiandrosterone (DHEA) is the androgen produced by the adrenal in largest

quantity. The physiological roles of DHEA and of its sulfate (DHEAS), which is thought to be produced exclusively by the adrenal cortex (Vermeulen, 1983), are unknown. They are considered weak androgens, but also appear to have estrogenic properties (Seymour-Munn and Adams, 1983). It has been proposed that DHEA and DHEAS protect against breast cancer in premenopausal women, but increase breast cancer risk in postmenopausal women (Bulbrook, 1971; Seymour-Munn and Adams,1983). These conflicting actions could be reconciled by a recent hypothesis: in premenopausal women, DHEA would have an antiestrogenic effect by binding competitively to estrogen receptors, whereas, in postmenopausal women, DHEA would bind to vacant estrogen receptors and enhance estradiol-like effects, thereby stimulating tumor growth (Ebeling and Koivisto, 1994).

Results from case-control studies of DHEA and DHEAS conducted in postmenopausal women have been mixed (Zumoff et al., 1981; Jones and James, 1987; Bernstein et al., 1990; Secreto et al., 1991). The three prospective cohort studies which examined the relationship of testosterone with breast cancer risk in postmenopausal women also measured DHEAS: one study (Barrett-Connor et al., 1990) reported no association whereas the two others (Gordon et al., 1990; Berrino, 1996) observed a non-significant positive association.

We report here on the relation between postmenopausal serum levels of testosterone and DHEAS and subsequent risk of breast cancer in a case-control study nested within a prospective cohort, the New York University (NYU) Women's Health Study. We previously reported a positive association between postmenopausal serum fractions of bioavailable estrogens and risk of breast cancer in this study population (Toniolo et al., 1995). A specific objective of our analysis was to examine whether serum levels of androgens have an effect on breast cancer risk independent of their influence on serum levels and biological availability of estrogens.

2 Methods

The NYU Women's Health Study cohort

Between March 1985 and June 1991, the NYU Women's Health Study enrolled a cohort of 14,275 women, 34 to 65 years old, at the Guttman Breast Diagnostic Institute, a breast cancer screening center in New York City. Details concerning subject recruitment have been published elsewhere (Toniolo et al., 1991; Toniolo et al., 1995). The current report is limited to the 7054 cohort members who were postmenopausal at the time of enrollment. Participants were classified as postmenopausal if they reported: (a) no menstrual cycles during the preceding 6 months, or (b) a total bilateral oophorectomy, or (c) a hysterectomy without complete oophorectomy prior to natural menopause and were 52 years of age or older. Cohort members donated 30 mL of blood and completed a self-administered questionnaire at enrollment. Blood was drawn prior to breast examination, between 9 A.M. and 3 P.M. in nonfasting women. After centrifugation, serum samples were immediately stored at $-80°C$ for subsequent biochemical analyses. Women who had taken hormonal medications in the 6 months preceding their visit were not eligible.

Nested case-control study

Cases of breast adenocarcinoma were identified primarily through active follow-up of the cohort and were confirmed by review of individual clinical and pathology records (Toniolo et al., 1995). For each case diagnosed in a woman who was postmenopausal at enrollment, two controls were selected at random from the risk set of women who were alive and free of disease at the time of diagnosis of the case, and who matched the case on age at enrollment ($\pm$6 months), date of initial blood donation ($\pm$3 months) and menopausal status. As of October 1991, 130 members of the postmenopausal cohort had been identified who had

received a diagnosis of breast cancer prior to January 1, 1991. Serum assays of follicle-stimulating hormone (FSH) were conducted to confirm the postmenopausal status of all the cases and their selected controls: three controls, who had reported the absence of menstrual cycles in the six months prior to enrollment, had FSH levels below17.5 IU/L, which was less than the minimal level compatible with postmenopausal status for our assay. They were nonetheless included in the analysis, since excluding them did not materially affect risk estimates. Estrogen assays (total estradiol, percent estradiol free, and percent estradiol bound to SHBG), were performed for all matched sets. For logistical reasons, androgen assays were carried out in a subset of 118 matched sets. Excluded from the analyses reported here are thirty-three matched sets for whom diagnosis of the case occurred 6 months or less after blood donation, 6 controls who reported treatment with corticosteroids in the 6 months prior to blood donation, and 1 control whose estrogen assays were done on a different day than the matching case. As a result, 85 cases (83 invasive and 2 noninvasive intraductal) and 163 controls are included in the present report.

Laboratory methods

For androgen assays, serum samples that had not been previously defrosted were shipped in dry ice to the Netherlands Cancer Institute and analyzed in two batches. Samples from a case and her matched controls were always analyzed in the same batch. All assays were performed in duplicate with the laboratory personnel blinded to the case or control status of the samples. Reference sera were included for each assay in several places within each batch.

Total testosterone was measured by a solid-phase radioimmunoassay (Coat-A-Count; Diagnostic Products Corp., Los Angeles, CA.) not requiring extraction or chromatography. The mean intra-assay coefficient of variation in the range of measurement was 6.2%. The inter-assay coefficients of variation were respectively 11% at 1.67, 2.2% at 10.00 and 7.0% at 22.05 nmol/L.

DHEAS was measured directly in diluted serum as we have reported previously (Bruning et al., 1984) using an antiserum against DHEA which showed a 42% crossreactivity with DHEAS. As DHEA was present in serum in concentrations at least 10 times lower than DHEAS, it had a negligible influence on the DHEAS values, which were read from a DHEAS standard plot. The mean intra-assay coefficient of variation in the range of measurement was 3.3%. The inter-assay coefficients of variation were respectively 10% at 1.68, 9% at 2.97, 10% at 5.51 and 8% at 15.18 ($\mu mol/L$).

Total estradiol was measured by standard radioimmunoassay (Pantex, Inc., Santa Monica, Calif.). Percent estradiol bound to SHBG and percent estradiol free were measured with a concanavalin A-Sepharose binding and an ultrafiltration method, respectively, as reported previously (Toniolo et al., 1995).

Statistical methods

When treated as continuous, total estradiol, testosterone and DHEAS were loge-transformed to reduce departures from the normal distribution. The paired t-test was used to compare hormone levels of the cases to the mean hormone levels of their matched controls.

To compute odds ratios, hormonal measurements were categorized into quartiles, using the frequency distribution of the cases and the controls combined. Since the androgen assays were performed in two batches, quartile cut-points were calculated separately for each batch. The weighted averages of the cut-points are reported in the tables.

The data were analyzed using conditional logistic regression (Breslow and Day, 1980). Odds ratios were computed relative to the lowest quartile. Regression analyses were also performed on the continuous hormonal variables. Likelihood ratio tests were used to assess the statistical significance of overall associations, linear trends and deviations from linearity. All p-values are two-sided.

One objective of the analysis was to examine concurrently the effect of androgens and

estrogens. Therefore, we report on the effect of adding androgen variables to models containing estrogen variables, and vice-versa. When adding estrogen variables to models containing androgen variables, percent SHBG-bound estrogen was entered first because it was the estrogen variable most strongly associated with breast cancer risk in multivariate models (Toniolo et al., 1995).

Hormone levels in this study were assessed from a single blood donation. For some hormones, however, a single measurement may not provide a reliable estimate of a woman's long-term average level, the exposure of interest, because of intrinsic fluctuations in the hormone over time and laboratory measurement error. In addition, different hormones are measured with varying amounts of error. For example, the reliability coefficients of total estradiol, percent estradiol bound to SHBG, and DHEAS, were estimated to be 0.51, 0.94, and 0.75, respectively, in our study population (Toniolo et al., 1994). The reliability of testosterone was not assessed in our study but estimates from the literature range from 0.74 (Micheli et al., 1991) to 0.88 (Hankinson et al., 1995). We were concerned that these differences might distort our results regarding the relative importance of the hormones. We therefore applied the method of Armstrong et al. (1989) for correcting logistic regression parameter estimates of continuous variables for measurement error in case-control data. For total estradiol, percent estradiol bound to SHBG, and DHEAS, we used within-subject variances which we had previously estimated (Toniolo et al., 1994). For testosterone, we used the within-subject variance estimate provided by Hankinson et al. (personal communication, 1996). We assumed that the different hormonal variables had independent measurement errors.

We examined the effect of Quetelet index (weight (kg) / height (m^2)) on the androgen-breast cancer associations, since the rate of conversion of androgens to estrogens increases with Quetelet index (Siiteri et al., 1973), and since the known positive association of Quetelet index with risk of breast cancer was confirmed in our data (Toniolo et al.,1995). The effect of other known risk factors (age at menarche, parity, age at first full-term pregnancy, age

at menopause, history of breast cancer in a first-degree relative, history of a benign breast condition, history of total oophorectomy, lifetime months of lactation and smoking history) on the androgen-breast cancer associations was also examined in multivariate conditional logistic analyses. The inclusion of covariates other than Quetelet index in the statistical analyses did not materially affect the results and are therefore not presented. In addition, the exclusion from the analysis of the 6 cases and 13 controls who had a total oophorectomy prior to enrollment in the study had no material impact on the results (data not shown). Results are therefore presented including these patients.

3 Results

Some characteristics of the study group are given in table 1. The median age at diagnosis of breast cancer was 61.6 years and the median duration between blood donation and diagnosis was 2.7 years (range 0.5 to 5.5 years). Known breast cancer risk factors had a similar distribution in this group as in the larger group on which estrogen assays were carried out (Toniolo et al., 1995). There were no appreciable differences between cases and controls in age at menarche, parity, age at menopause and history of prior oophorectomy. Delayed first full-term pregnancy, history of breast cancer in at least one first degree relative 45 years old or younger and history of a benign breast condition were associated with a non-significant increase in risk of breast cancer, while a history of breast-feeding was associated with a non-significant protective effect. The median weight and median Quetelet index were significantly higher in cases than in controls.

Table 2 shows the geometric mean levels of testosterone and DHEAS for cases and controls. The mean testosterone level was 21% higher in cases than in controls ($p < 0.01$) and the mean DHEAS level was 20% higher ($p = 0.10$).

Table 3 reports odds ratios for the association between breast cancer and serum levels of testosterone, total estradiol and percent estradiol bound to SHBG. In univariate analyses,

odds ratios showed a significant increase (p=0.02, test for trend) in risk of breast cancer with increasing levels of testosterone: the odds ratios for the second, third and fourth quartiles relative to the lowest quartile, were 2.4 (95% CI, 1.0-5.6), 3.5 (95% CI, 1.4-8.4) and 2.7 (95% CI, 1.1-6.8), respectively. However, adjusting for percent SHBG-bound estradiol, which was the estrogen variable most strongly associated with breast cancer risk, reduced the odds ratios and removed the significant trend. The odds ratios were 1.5 (95% CI, 0.6-3.7), 2.0 (95% CI, 0.7-5.2) and 1.3 (95% CI, 0.5-3.7) for second, third and fourth quartiles respectively. Adding total estradiol to the model including testosterone and percent SHBG-bound estradiol did not significantly improve the fit of the model, although it further reduced the odds ratios to 1.4 (95% CI, 0.6-3.5), 1.8 (95% CI, 0.7-5.0) and 1.2 (95% CI, 0.4-3.5) respectively. Adding percent free estradiol or Quetelet index to the model containing testosterone, estradiol and percent estradiol bound to SHBG did not materially affect the odds ratios (data not shown). A strong positive association between breast cancer risk and increasing levels of total estradiol was also present in univariate analysis. This association remained significant after adjusting for testosterone levels, although the odds ratios and the corresponding p-value were somewhat reduced. The protective effect associated with increasing percentage of SHBG-bound estradiol was hardly affected by adjustment for testosterone levels. In the model including the 3 hormonal variables, only the percentage of estradiol bound to SHBG remained significant. Analysis on continuous variables showed similar results.

Results of the analyses correcting for measurement error were similar to results of the uncorrected analyses with respect to the relative strength of the associations of the hormonal variables with breast cancer risk: the positive association of testosterone was weakened and no longer significant after adjusting for percent SHBG-bound estradiol, whereas the positive association of total estradiol became only marginally significant and the negative association of percent SHBG-bound estradiol remained highly significant after adjusting for testosterone. In the model including the three variables, only percent SHBG-bound estradiol remained significant.

Table 4 reports odds ratios for the association between breast cancer risk and increasing levels of DHEAS. In unadjusted analyses, although the odds ratio in the highest quartile was slightly elevated (1.6, 95% C.I: 0.7-3.5), there was no trend of increasing risk of breast cancer with increasing levels of DHEAS. The inclusion of estrogen variables or of Quetelet index did not result in a significant trend. The odds ratios for the association between breast cancer risk and DHEAS are shown adjusting for percent SHBG-bound estradiol and total estradiol. Inclusion of DHEAS in models containing estrogen variables did not materially affect the associations between estrogen variables and breast cancer risk (data not shown). Correcting for measurement error in the hormonal variables did not alter the results.

Finally, analyses were conducted using only the 56 matched sets with at least two years between blood donation and diagnosis of the case. The results were similar to the results of analyses conducted in the larger group, both for testosterone and DHEAS (data not shown).

Table 5 reports the Spearman correlation coefficients for hormone levels and Quetelet index, by case-control status. Note that testosterone was correlated positively with total estradiol ($r_S = 0.23$ in cases and 0.27 in controls) and negatively with percent estradiol bound to SHBG ($r_S = -0.27$ in cases and -0.33 in controls).

4 Conclusions

In unadjusted analyses (except for matching variables), we observed a statistically significant trend of increasing risk of breast cancer with increasing serum levels of testosterone in postmenopausal women. Since all cases were diagnosed at least 6 months after blood donation (median 2.7 years) and since a similar trend was observed when the analysis was limited to the two-thirds of the cases diagnosed at least two years after blood donation, it seems unlikely that the higher levels of testosterone observed in women who subsequently developed the disease, compared with controls, resulted from the presence of tumors.

Three previous prospective studies have examined the association between serum levels of

testosterone and breast cancer risk in postmenopausal women. No association was observed in the Rancho Bernardo, California, study, in which the age-adjusted mean testosterone level was 258 pg/ml in 15 cases diagnosed at least 1 year after blood donation and 261 pg/ml in 400 noncases (Garland et al., 1992). However, results from the two other prospective studies are consistent with ours. In the Washington County, Maryland, study, serum levels were 11% higher in 39 cases (mean 304 pg/ml) than in 155 controls (mean 274 pg/ml), although this difference was not statistically significant (Wysowski et al., 1987). Finally, in 24 cases diagnosed during the first 3.5 years of follow-up of a cohort of 4040 postmenopausal women from northern Italy, the risk ratios for breast cancer associated with the second and third tertiles of testosterone were 4.8 (95% CI 0.9-25.1) and 7.0 (95% CI 1.4-36.4), respectively (p for trend = 0.026, Berrino et al., 1996).

We recently reported a positive association between bioavailable estrogens and subsequent risk of breast cancer in a slightly larger group of postmenopausal women from the NYU Women's Health Study (Toniolo et al., 1995). An objective of the present analysis was to examine whether androgens had an effect on breast cancer risk that was independent of their influence on serum levels and biological availability of estrogen. Results showed that, after including estrogen variables (percent SHBG-bound estradiol and total estradiol) in our statistical model, the odds ratios associated with higher levels of testosterone were considerably reduced, and there was no longer a significant trend. A similar result was recently observed in the re-analysis of a population-based case-control study conducted in Sweden (Adami et al. 1979; Lipworth et al., 1996). Whereas in univariate analysis, a significant positive association was found between testosterone and breast cancer risk, the association disappeared after controlling for estrone (and androstenedione). On the other hand, Berrino et al. (1996) did not observe a reduction of the association between breast cancer risk and levels of testosterone when adjusting for total estradiol. However, multivariate analysis was hampered by the small sample size of the study (24 cases).

We were concerned about the impact of measurement error in the hormonal variables on

our results. It is well known that in the absence of confounders, non-differential measurement error in an exposure variable will result in an attenuation of the true exposure/disease relationship. When several variables are measured with error, however, the associations of these variables with disease in a multivariate model may be weaker or stronger than the true associations (Armstrong et al., 1990). It is reassuring that, in our analysis, correcting for measurement error did not affect the relative strength of the associations of the hormonal variables with risk of breast cancer.

Our results are consistent with the hypothesis that testosterone has an indirect effect on breast cancer risk, through its association with estrogen levels. The fact that percent SHBG-bound estradiol was the estrogen variable which caused the greatest reduction in the testosterone-breast cancer odds ratios suggests that the effect of testosterone on the bioavailability of estrogens may be more important than its role as a precursor of estrogens. An increase in the serum levels of testosterone could lead to a decrease in the percentage of estradiol bound to SHBG since testosterone binds to SHBG with greater affinity than estradiol. However, the modeling studies performed by Dunn et al. (1981) as well as in vitro experiments (Bonfrer et al., 1989) indicate that higher concentrations of testosterone would be required to observe such an effect. Inhibition of the hepatic secretion of SHBG by testosterone could also result in a decrease in percent SHBG-bound estradiol, since small changes in SHBG concentration can produce an important reduction in the percentage of hormone bound to this protein (Selby, 1990). In support of this hypothesis a moderate negative correlation between testosterone and SHBG was reported by some (Haffner et al., 1995; Lonning et al., 1995; Maggino et al., 1993) although not all (Lipworth, 1996) studies.

A limitation of our study is that only total testosterone was measured. The free and albumin-bound hormone fractions might be more relevant biologically since these fractions are thought to diffuse readily into the cells (Pardridge et al., 1981). Indeed, with regard to estrogen, the variable most strongly related to risk of breast cancer was the percent SHBG-bound estradiol which had a protective effect. Thus, we cannot exclude the possibility that

the free and albumin-bound fractions of testosterone might have an independent effect on breast cancer risk.

The lack of an association between DHEAS and breast cancer observed here is consistent with the results of the previous prospective studies which examined the role of this hormone in postmenopausal women. Barrett-Connor et al. (1990) measured DHEAS levels in a cohort of 534 women, 50 to 79 years old, among whom 21 subsequently developed breast cancer, and reported no difference between cases and non-cases. In a case-control study nested within a cohort of approximately 13,000 female residents of Washington County, Maryland, Gordon et al. (1990) reported that serum levels of DHEA were significantly higher in 30 postmenopausal women who developed breast cancer 9 years or more after blood donation than in 59 matched controls. However, no statistically significant difference in DHEAS levels was observed, although serum levels of DHEAS were slightly higher in the women who developed breast cancer than in the controls. Finally, an increase in the odds ratios for breast cancer was observed with increasing serum levels of DHEAS in the ORDET study but this trend was not statistically significant (Berrino et al., 1996). Overall, there is little epidemiologic evidence that DHEAS plays an important role in breast cancer development in postmenopausal women.

In conclusion, elevated serum levels of testosterone were found to be associated with subsequent risk of breast cancer in postmenopausal women. However, this association was considerably reduced and no longer significant after taking into account the effect of serum estrogen levels on breast cancer risk, suggesting that androgens act through their influence on the availability of estrogens via SHBG binding and/or as precursors of estrogens. There was no evidence that the adrenal androgen DHEAS plays a role in breast cancer development in our study. In light of these results, additional research to identify factors influencing testosterone levels in healthy postmenopausal women would be of interest. Among life-style factors such as smoking, obesity, diet, alcohol consumption and exercise, only obesity has been found to be marginally associated with higher levels of testosterone (Cauley et al., 1989;

Newcomb et al., 1995).

ACKNOWLEDGMENTS

We thank the patients, staff and management of the Guttman Breast Diagnostic Institute for their collaboration. We thank Anahita Bassiri, Elizabeth Clancy, Jo-Ann Cutrone, Frances Mastrota, Lynne Quinones, Joan Szymczak and Cristina Trigo for their dedication to the study and their skills in data collection and management. We thank Dorothe Linders and Tiny Korse from the Netherlands Cancer Institute for their excellent technical assistance.

REFERENCES

Adami HO, Johansson EDB, Vegelius J, Victor A: Serum concentrations of estrone, androstenedione and sex-hormone-binding globulin in postmenopausal women with breast cancer and in age-matched controls. Upsala J Med Sci 1979; 84: 259-274.

Armstrong BG, Whittemore AS, Howe GR: Analysis of case-control data with covariate measurement error: Application to diet and colon cancer. Stat Med 1989; 8: 1151-1163.

Armstrong BG: The effects of measurement errors on relative risk regressions. Am J Epidemiol 1990; 132: 1176-1184.

Barret-Connor E, Friedlander NJ, Khaw KT: Dehydroepiandrosterone sulfate and breast cancer risk. Cancer Res 1990; 50: 6571-6574.

Bernstein L, Ross RK, Pike MC, Brown JB, Henderson BE: Hormone levels in older women: a study of postmenopausal breast cancer patients and healthy population controls. Br J Cancer 1990; 61: 298-302.

Bernstein L and Ross RK: Endogenous hormones and breast cancer risk. Epidemiologic Rev 1993; 15: 48-65.

Berrino F, Muti P, Micheli A, Bolelli G, Krogh V, Sciajno R, Pisani P, Panico S, Secreto G: Serum sex hormone levels after menopause and subsequent breast cancer. J Natl Cancer Inst 1996; 88: 291-296.

Bonfrer JMG, Bruning PF and Nooijen WJ: A simple method for the measurement of the

steroid fraction bound to sex-hormone bound globulin in serum. J Steroid Biochem 1989; 33: 227-231.

Breslow NE, Day NE: Statistical methods in cancer research, vol 1. The analysis of case-control studies. IARC Scientific Publications No. 32. Lyon, 1980.

Bruning PF, Bonfrer JMG, De Jong-Bakker M, Nooyen W: Influence of ACTH on aminogluthimidine induced reduction of plasma steroids in postmenopausal breast cancer. J Steroid Biochem 1984; 21:293-298.

Bulbrook RD, Hayward JL, Spicer CC: Relation between urinary androgen and corticosteroid excretion and subsequent breast cancer. Lancet 1971; 2: 395-397.

Cauley JA, Gutai JP, Kuller LH, LeDonne D, Powell JG: The epidemiology of serum sex hormones in postmenopausal women. Am J Epidemiol 1989; 129: 1120-1131.

Dunn JF, Nisula BC and Rodbard D: Transport of steroid hormones: binding of 21 endogenous steroids to both testosterone-binding globulin and cortico-steroid binding globulin in human plasma. J Clin Endocrinol Metab 1981; 53: 58-68.

Ebeling P, Koivisto VA: Physiological importance of dehydroepiandrosterone. Lancet 1994; 343: 1479-1481.

Garland CF, Friedlander NJ, Barrett-Connor E, Khaw KT: Sex hormones and postmenopausal breast cancer: A prospective study in an adult community. Am J Epidemiol 1992; 135: 1220-1230.

Gordon GB, Bush TL, Helzlouer KJ, Comstock GW: Relationship of serum levels of dehydroepiandrosterone and dehydroepiandrosterone sulfate to the risk of developing postmenopausal breast cancer. Cancer Res 1990; 50: 3859-3862.

Grattarola R, Secreto G, Recchione C, Castellini W: Androgens in breast cancer. II. Endometrial adenocarcinoma and breast cancer in married postmenopausal women. Am J Obstet Gynecol 1974; 118: 173-178.

Haffner SM, Newcomb PA, Marcus PM, Klein BEK, Klein R: Relation of sex hormones and dehydroepiandrosterone sulfate (DHEA- SO4) to cardiovascular risk factors in postmenopausal women. Am J Epidemiol 1995; 142: 925-934.

Hankinson SE, Manson JE, Spiegelman D, Willett WC, Longcope C and Speizer FE: Reproducibility of plasma hormone levels in postmenopausal women over a 2-3 year period. Cancer Epidemiol Biomarkers Prev 1995; 4:649-654.

Henderson BE, Ross RK and Pike MC: Hormonal chemoprevention of cancer in women. Science, 1993; 259: 633-638.

Hill P, Garbaczewski L, Kasumi F: Plasma testosterone and breast cancer. Eur J Cancer Clin Oncol 1985; 21: 1265-1266.

Jones DL, James VHT: Determination of dehydroepiandrosterone and dehydroepiandrosterone sulphate in blood and tissue: studies of normal women and women with breast or endometrial cancer. J Steroid Biochem 1987; 26: 151-159.

Lipworth L, Adami HO, Trichopoulos D, Carlstrom K, Mantzoros C: Serum steroid hormone

levels, sex hormone-binding globulin and body mass index in the etiology of postmenopausal breast cancer. Epidemiology 1996; 7:96-100.

Lonning PE, Helle SI, Johannessen DC, Adlercreutz H, Lien EA, Tally M, Ekse D, Fotsis T, Anker GB, Hall K: Relations between sex hormones, sex hormone binding globulin, insulin-like growth factor binding protein-1 in postmenopausal breast cancer patients. Clinical Endocrin 1995; 42: 23-30.

McFadyen IJ, Forrest APM, Prescott RJ, Groom GV, Golder MP, Fahmy DR: Circulating hormone concentrations in women with breast cancer. Lancet 1976; i: 1100-1102.

Maggino T, Pirrone F, Velluti F, Bucciante G: The role of the endocrine factors and obesity in hormone-dependent gynecological neoplasias. Eur J Gynaec Oncol 1993; vol XIV, pp 119-126.

Malarkey WB, Schroeder LL, Stevens VC, James AG, Lanese RR: Twenty-four hour preoperative endocrine profiles in women with benign and malignant breast disease. Cancer Res. 1977; 37: 4655-4659.

Micheli A, Muti P, Pisani P, Secreto G, Recchione C, Totis A, Fissi R, Cavalleri A, Panico S and Berrino F: Repeated serum and urinary androgen measurements in premenopausal and postmenopausal women. J Clin Epidemiol 1991; 44: 1055-1061.

Newcomb PA, Klein R, Klein BEK, Haffner S, Mares-Perlman J, Cruikshanks KJ, Marcus PM: Association of dietary and life-style factors with sex hormones in postmenopausal women. Epidemiology 1995; 6: 318-321.

Pardridge WM: Transport of protein-bound hormones into tissues in vivo. Endocrine Reviews 1981; 2: 103-121.

Secreto G, Recchione C, Cavalleri A, Miraglia M, Dati V: Circulating levels of testosterone, 17(-oestradiol, luteinising hormone and prolactin in postmenopausal breast cancer patients. Br J Cancer 1983, 47: 269-275.

Secreto G, Toniolo P, Berrino F, Recchione C, Cavalleri A, Pisani P, Totis A, Fariselli G, Di Pietro S: Serum and urinary androgens and risk of breast cancer in postmenopausal women. Cancer Res 1991; 51: 2572-2576.

Secreto G, Zumoff B: Abnormal production of androgens in women with breast cancer (Review). Anticancer Res 1994; 14: 2113-2117.

Selby C: Sex hormone binding globulin: Origin, function and clinical significance. Ann Clin Biochem 1990; 27: 532-541.

Seymour-Munn K, Adams J: Estrogenic effects of 5-androstene-3 beta, 17 beta-diol at physiological concentrations and its possible implication in the etiology of breast cancer. Endocrinol 1983; 112: 486-491.

Siiteri PK, MacDonald PC: Role of extraglandular estrogen in human endocrinology. In: Geiger SR, Astwood EB, Greep RO, eds. Handbook of physiology. Washington DC: American Physiological Society, 1973: 615-629.

Siiteri PK, Hammond GL, Nisker JA: Increased availability of serum estrogens in breast cancer: a new hypothesis. In: Banbury Report 8 (Pike MC, Siiteri PK, Welsh CW, eds).

Cold Spring Harbor Laboratory, 1981, pp 87-101.

Toniolo PG, Pasternack BS, Shore RE, Sonnenschein E, Koenig KL, Rosenberg C, Strax P, Strax S: Endogenous hormones and breast cancer: A prospective cohort study. Breast Cancer Res Treat 1991; 18: S23-S26.

Toniolo P, Koenig KL, Pasternack BS, Banerjee S, Rosenberg C, Shore RE, Strax P and Levitz M: Reliability of measurements of total, protein-bound, and unbound estradiol in serum. Cancer Epidemiol Biomarkers Prev 1994; 3: 47-50.

Toniolo P, Levitz M, Zeleniuch-Jacquotte A, Banerjee S, Koenig KL, Shore RE, Strax P, Pasternack BS: A prospective study of endogenous estrogens and breast cancer in post-menopausal women. J Natl Cancer Inst 1995; 87: 190-197.

Vermeulen A: The androgens. In: Gray CH, James VHT, eds. Hormones in blood, 3rd edition, vol 4. New York: Academic Press, Inc, 1983: 349-361.

Wysowski DK, Comstock GW, Helsing KJ, Lau HL: Sex hormone levels in serum in relation to the development of breast cancer. Am J Epidemiol 1987; 125: 791-799.

Zumoff B, Levin J, Rosenfeld RS, Markham M, Strain GW, Fukushima DK: Abnormal 24-hr mean plasma concentrations of dehydroepiandrosterone and dehydroepiandrosterone sulfate in women with primary operable breast cancer. Cancer Res 1981; 41: 3360-3363.

Table 1. Study group characteristics.

	Median (range) or percent	
	Cases (n=85)	Controls (n=163)
Age at blood donation	59.2 (48.9-65.4)	59.1 (48.9-64.9)
Age at diagnosis	61.6 (52.2-68.6)	
Age at menarche	**13 (9-16)**	**13 (10-17)**
No. of full-term pregnancies		
0	**24.7%**	**23.9%**
1	**17.6%**	**13.5%**
>1	**57.6%**	**62.6%**
Age at first full-term pregnancy	**25 (16-41)**	**24 (16-43)**
Ever breast-feeding	**20.8%**	**28.2%**
Age at menopause	**51.7 (31.6-57.2)**	**50.9 (24.9-58.6)**
Breast cancer in first degree relative		
< 45 years old	**8.2%**	**3.7%**
Prior benign breast condition	**57.7%**	**46.7%**
Prior bilateral oophorectomy	**7.0%**	**8.0%**
Height, cm	**162.6 (149.9-177.8)**	**162.6 (147.3-177.8)**
Weight***, kg	70.3 (47.6-122.5)	62.6 (45.4-124.7)
Quetelet's index***, kg/m^2	26.1 (19.9-43.6)	24.0 (17.7-44.4)

*** $p < 0.001$, paired t-test

Table 2. Geometric mean, geometric standard deviation and range of serum levels of testosterone and DHEAS in breast cancer patients diagnosed at least 6 months after blood donation, and their individually-matched controls.

	Cases (n=85)	Controls (n=163)
Testosterone (nmol/L)		
Mean** (S.D.)	1.05 (1.79)	0.87 (1.89)
Range	0.20-3.96	0.14-5.96
DHEAS (μmol/L)		
Mean* (S.D.)	2.36 (2.37)	1.96 (2.26)
Range	0.22-14.60	0.12-10.43

* $p = 0.10$, paired t-test

** $p < 0.01$, paired t-test

Table 3. Odds ratios for the association between breast cancer risk and serum levels of testosterone, total estradiol and percent SHBG-bound estradiol.

Hormonal variable by quartiles	OR[1] (95% CI)	OR[2] (95% CI)	OR[3] (95% CI)
Testosterone[4]			
1	1.0	1.0	1.0
2	2.4 (1.0-5.6)	1.5 (0.6-3.7)	1.4 (0.6-3.5)
3	3.5 (1.4-8.4)	2.0 (0.7-5.2)	1.8 (0.7-5.0)
4	2.7 (1.1-6.8)	1.3 (0.5-3.7)	1.2 (0.4-3.5)
p for trend	*	N.S.	N.S.
Total estradiol[5]			
1	1.0	1.0	1.0
2	2.0 (0.8-5.3)	1.8 (0.7-4.8)	1.7 (0.6 - 4.7)
3	4.3 (1.8-10.4)	3.6 (1.4-9.0)	2.6 (1.0 - 6.8)
4	3.8 (1.5-10.3)	2.9 (1.0-8.3)	1.6 (0.5 - 5.8)
p for trend	***	*	N.S.
% SHBG-bound estradiol[6]			
1	1.0	1.0	1.0
2	0.43 (0.19-0.98)	0.44 (0.19-1.01)	0.44 (0.19-1.05)
3	0.19 (0.07-0.49)	0.20 (0.07-0.56)	0.21 (0.07-0.59)
4	0.05 (0.01-0.17)	0.05 (0.01-0.19)	0.05 (0.01-0.21)
p for trend	***	***	***

* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

[1] Unadjusted, except for matching factors (age and serum storage time).

[2] For testosterone, ORs are adjusted for % SHBG-bound estradiol; for total estradiol and % SHBG-bound estradiol, ORs are adjusted for testosterone.

[3] Adjusted for other hormonal variables in the table.

[4] The cut-points defining quartiles of testosterone were 0.73, 1.02 and 1.45 nmol/mL.

[5] The cut-points defining quartiles of total estradiol were 20, 30 and 45 pg/mL.

[6] The cut-points defining quartiles of percent SHBG-bound estradiol were 34.4, 43.6 and 51.3%.

Table 4. Odds ratios for the association between breast cancer risk and serum levels of DHEAS.

Quartiles of DHEAS[1]	OR[2] (95% CI)	OR[3] (95% CI)
1	1.0	1.0
2	0.7 (0.3-1.5)	0.3 (0.1-0.9)
3	1.0 (0.5-2.1)	0.5 (0.2-1.3)
4	1.6 (0.7-3.5)	0.9 (0.4-2.3)
p for trend	**N.S.**	**N.S**

[1] The cutpoints defining quartiles of DHEAS were 1.33, 2.38 and 3.58 μmol/L.

[2] Unadjusted, except for matching factors (age and serum storage time).

[3] Adjusted for percent SHBG-bound estradiol and total estradiol.

Table 5. Spearman correlation coefficients for androgen and estrogen levels and Quetelet index.

Controls (n = 163)

	DHEAS	Total estradiol	% SHBG-bound estradiol	% free estradiol	Quetelet index
Testosterone	0.35***	0.27***	-0.33***	0.25**	0.30***
DHEAS		0.23**	-0.27***	0.25**	0.05
Total estradiol			-0.48***	0.45***	0.43***
% SHBG-bound estradiol				-0.72***	-0.52***
% free estradiol					0.48***

Cases (n = 85)

	DHEAS	Total estradiol	% SHBG-bound estradiol	% free estradiol	Quetelet index
Testosterone	0.38***	0.23*	-0.27*	0.12	0.11
DHEAS		0.28*	-0.37***	0.24*	-0.06
Total estradiol			-0.47***	0.19	0.38***
% SHBG-bound estradiol				-0.56***	-0.42***
% free estradiol					0.29**

* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

Chapter III

Sample Size and Study Design Considerations for Half-Life Studies

1 Introduction

The accumulation of PCBs (polychlorinated biphenyls) and DDE (1,1 dichloro-2,2-bis(p-chlorophenyl) ethylene) residues, and other environmental contaminants in the body may potentially have adverse health effects. Individuals who are able to clear these toxic compounds from the body at a faster rate, and thus have shorter half-lives, may be at lower risk of diseases associated with the toxins. Thus, in order to fully elucidate the role of environmental contaminants in the development of disease, their rates of persistence in the body must be accurately quantified.

Previous studies estimating the half-life of PCBs have yielded inconsistent results, however. Reported estimates of half-life range from .5 months to 17 years for PCB mixtures. (Yakushiji et al., 1984; Phillips et al., 1989; Elo et al 1985; Lawton et al., 1985). For specific PCB components, half-lives have been estimated to be from less than 1 year to about 30 years (Yakushiji et al, 1984; Chen et al, 1982). Similarly, data on the half-life of DDE are variable and limited.

The lack of consistency among study estimates of half-life may be largely due to the small sample sizes and limited number of repeated measurements per subject utilized in these studies. For example, Chen et al (1982) examined the rates of elimination of PCBs from the blood of PCB-poisoned subjects in Taiwan using two to three serial blood samples from 17 individuals taken over a period of 6-14 months. Similarly, Steele et al (1986) calculated the half-life of PCBs using two measurements of PCB concentrations made 7 years apart.

Phillips (1989) investigated how analytical (laboratory) error and the time interval between measurements affect the variability and possible bias in estimates of half-life calculated from two measurements. Results indicate that half-life estimates based on only two measurements become increasingly variable at shorter time intervals between measurements and at higher levels of analytical error.

The precision of half-life estimates, however, is not only dependent on the magnitude of

analytical error and the time interval between measurements, but also on the number of repeated measurements utilized in the estimation procedure. Given laboratory cost contraints, time constraints, and other limitations on the physical resources of a study on half-life, investigators must decide where to allocate the resources in order to obtain the most precise estimate of half-life.

Issues of sample size and study design for estimating subject-specific, as well as population half-lives of environmental contaminants have not been formally addressed in the environmental and epidemiologic literature. The objectives of this paper are to provide useful guidelines for choosing the number of repeats and the optimal time interval between repeats needed for estimating an individual's half-life with a given level of precision, while minimizing the cost of the study. In addition, sample size and power considerations for studies comparing the population half-lives between two groups will be investigated. An example is presented using data from a study on PCBs and breast cancer.

2 Methods

For most environmental toxins, the rate of elimination from the body may be described by the following one-compartment exponential decay model:

$$C(t) = C_0 e^{-\lambda t}, \tag{1}$$

where C(t) is the concentration of the toxin at time t, C_0 denotes the initial concentration, and λ is the rate constant. The half-life, $t_{1/2}$, which is the time after which the level of toxin is reduced to half its original value, is equal to $\ln(2)/\lambda$.

If both sides of (1) are log-transformed, then we have the linear relationship:

$$\ln\{C(t)\} = \ln(C_0) - \lambda t. \tag{2}$$

Thus, given $\mathbf{C(t)} = \{C(t_1), ..., C(t_k)\}$, the set of serial measurements of the toxin obtained on a subject at times, $\{t_1, ..., t_k\}$, the rate constant, λ, may be estimated from the slope

of the linear regression of $\ln\{\mathbf{C(t)}\}$ versus $\mathbf{t}$. The least-squares estimate of λ is equal to $\hat{\lambda} = \sum_{j=1}^{k}\{\ln\{C(t_j)\} - \bar{C}\}(t_j - \bar{t})/\sum_{j=1}^{k}(t_j - \bar{t})^2$, where $\bar{C}$ and $\bar{t}$ denote the average logarithm of the level of toxin and average time of measurement, respectively. The corresponding half-life for the subject may be estimated by $\hat{t}_{1/2} = \ln(2)/\hat{\lambda}$.

The sample size and study design issues associated with estimating the half-life will depend on whether the focus is on obtaining a precise estimate of an individual's half-life or a population half-life. The former would be of interest, for example, in studies exploring the relationship between an individual's rate of elimination of the toxin with a particular genetic characteristic. On the other hand, a precise estimate of a population half-life would be pertinent when the investigator is interested in comparing the average half-lives between two or more groups, such as diseased and non-diseased subjects.

Study Design for Estimating Individual Half-Lives

If the goal is to estimate individual half-lives with a certain level of precision, then clearly, the number of subjects to include in the study is not relevant. The frequency of measurement and duration of follow-up are the primary factors which will determine the precision of the individual's half-life estimate. This can be shown as follows.

The variance of $\hat{\lambda}$, the least-squares estimate of the rate parameter, is equal to $\sigma_e^2/\sum_{j=1}^{k}(t_j - \bar{t})^2$, where σ_e^2 is the variance of the deviation of the observed $\ln\{\mathbf{C(t)}\}$ from the value predicted by the regression line in (2). Then, using the Delta method (Cox and Hinkley, 1974), the variance of $\hat{t}_{1/2}$ is equal to,

$$V(\hat{t}_{1/2}) = \ln(2)^2 \left(\frac{1}{\lambda}\right)^4 \frac{\sigma_e^2}{\sum_{j=1}^{k}(t_j - \bar{t})^2}. \tag{3}$$

Let $\mathbf{t} = \{t_1, t_2, ..., t_k\}$ denote k equally spaced points in time, where the time interval between points is equal to I. Then the study duration, D, is equal to $I(k-1)$. Following the arguments in Schlesselman (1973), we can express $\sum_{j=1}^{k}(t_j - \bar{t})^2$ as a function of D and

k:

$$\sum_{j=1}^{k}(t_j - \bar{t})^2 = \frac{D^2 k(k+1)}{\{12(k-1)\}}.$$

It follows that the variance of $\hat{t}_{1/2}$ can be expressed as:

$$V(\hat{t}_{1/2}) = \ln(2)^2\left(\frac{1}{\lambda}\right)^4 \frac{\sigma_e^2\{12(k-1)\}}{D^2 k(k+1)} = \left(\frac{t_{1/2}}{\lambda D}\right)^2 \frac{\sigma_e^2\{12(k-1)\}}{k(k+1)}. \quad (4)$$

Thus, (4) describes how the precision of $\hat{t}_{1/2}$ is a function of the study duration, D, the number of repeated measurements on a subject, k, σ_e, and λ. For fixed values of the underlying rate parameter, λ, and σ_e, the variance of $\hat{t}_{1/2}$ is directly proportional to $\omega = \frac{ln(2)^2 12(k-1)}{D^2 k(k+1)}$. Schlesselman presented tables which show how the precision of a slope changes over different values of k and D. Table 1 describes analogous results for the precision of the half-life. Specifically, we calculated ω for various values of k and D. One can easily see how ω, and thus, the variance of the half-life, decreases as the number of repeats and the duration of study increases. The exception, however, is that for a fixed duration of study, obtaining 3 measurements does not result in additional precision compared with 2 measurements. (This is due to the algebraic result that the term $(k-1)/k(k+1)$ in (4) is the same for $k = 2$ or 3.) Furthermore, for large k, the variance of $\hat{t}^i_{1/2}$ is proportional to $1/(D^2 k)$. Thus, a unit increase in the duration of the study will result in greater precision of the half-life estimate than a unit increase in the number of repeated measurements. Finally, note that some combinations of k and D will yield the same level of precision. For example, 10 measurements obtained over 7 months result in the same precision as 7 measurements over 8 months, and 3 measurements over 10 months.

The choice between different pairs of (k, D) for estimating the half-life will depend upon the relative costs of each measurement and each time interval of follow-up (which may include staff salaries and other administrative costs). If the two costs are equivalent, then results from Table 1 suggest that resources should be directed toward extending the duration of the study, since this will result in larger gains in precision than will increasing the number of

measurements. When the costs of (k, D) differ, however, the allocation of resources which will result in the most precise estimate of $t_{1/2}$ is less clear.

For each subject, let $C = c_1 k + c_2 D$ equal the total cost of measuring the subject k times over a duration of D years, where c_1 denotes the cost of an individual measurement, and c_2 denotes the cost per year of follow-up. Assume that the goal is to estimate an individual's half-life with variance equal to V, while minimizing the total cost per study subject. If we make the simplifying assumption that for large k,

$$V \approx \ln(2)^2 \left(\frac{1}{\lambda}\right)^4 \frac{12\sigma_e^2}{D^2 k} = \left(\frac{t_{1/2}}{\lambda D}\right)^2 \frac{12\sigma_e^2}{k}, \tag{5}$$

then a Lagrange multiplier may be used to minimize C subject to the constraint in (5). After some algebraic manipulations, we have the result

$$k = \left\{\frac{\ln(2)^2 3 c_2^2 \sigma_e^2}{V c_1^2 \lambda^4}\right\}^{1/3} = \left\{\frac{3\sigma_e^2}{V}\left(\frac{c_2}{c_1}\right)^2 \left(\frac{t_{1/2}}{\lambda}\right)^2\right\}^{1/3}, \tag{6}$$

and

$$D = \left\{\frac{\ln(2)^2 24 c_1 \sigma_e^2}{V c_2 \lambda^4}\right\}^{1/3} = \left\{\frac{24\sigma_e^2}{V}\left(\frac{c_1}{c_2}\right) \left(\frac{t_{1/2}}{\lambda}\right)^2\right\}^{1/3}, \tag{7}$$

as the optimal values of k and D which will minimize the cost for a specified level of precision, V. As expected, the optimal k and D depend on c_2/c_1, the ratio of the cost per month of follow-up to the cost per measurement. As this ratio increases, the optimal design favors increasing the number of repeated measurements and decreasing the duration of follow-up. In order to calculate k and D from (6) and (7), respectively, values of λ and σ_e^2 must be assumed. Estimates may be obtained from the literature or preliminary studies.

The above result is valid only when k is large enough so that $(k-1)/(k+1) \approx 1$. When this assumption does not hold, closed form solutions are not available for calculating the optimal k and D, and iterative methods must be utilized. Investigators who are unfamiliar with iterative numerical techniques may need to consult a statistician.

Study Design for Estimating and Comparing Population Half-Lives

In the above discussion, it was assumed that the primary focus was on estimating the subject-specific half-lives. Thus, the size of the study population was not relevant. However, when the goal is to estimate the average half-life in a particular population, or to compare the half-lives in two different populations, then one needs to consider the number of subjects to include in the study, in addition to the frequency and duration of measurements.

Assume that the sample population is comprised of N subjects, and that each subject has a "true" rate parameter, λ_i, which is distributed with mean λ_P and variance, σ_s^2. Thus, λ_P can be interpreted as the underlying population rate parameter, and σ_s^2 is the variance in λ_i between individuals. Furthermore, assume that the frequency of measurement, study duration, and σ_e are the same for all subjects.

Given the estimated subject-specific half-lives: $\{\hat{t}^1_{1/2}, ..., \hat{t}^N_{1/2}\}$, the population half-life, $t^P_{1/2}$, may be estimated by: $\hat{t}^P_{1/2} = \{\hat{t}^1_{1/2} + ... + \hat{t}^N_{1/2}\}/N$. Using result (4) and the assumptions above, it can be shown that the variance of $\hat{t}^P_{1/2}$ is equal to

$$V(\hat{t}^P_{1/2}) = \ln(2)^2(\frac{1}{\lambda_P})^4 \left[\sigma_s^2 + \frac{\sigma_e^2\{12(k-1)\}}{D^2k(k+1)}\right]\frac{1}{N}. \tag{8}$$

Equation (8) can be used to determine the k, D, and N which will result in a certain level of precision in the population half-life estimate. One can see from the form of the equation that the precision of $\hat{t}^P_{1/2}$ improves as k, D, and N increase, and that increases in N will diminish both the contributions of σ_s^2 and σ_e^2 to the variance. Note that the variance is no longer directly proportional to a factor which is a function only of k, D and N. Thus, tables similar to Table 1 cannot be generated unless values for σ_e^2 and σ_s^2 are assumed. The use of (8) will be illustrated in the example.

Design issues for studies comparing the half-lives between two populations will now be considered. Let $t^1_{1/2}$ and $t^2_{1/2}$ denote the half-lives in the two populations. The null hypothesis is $H_0 : t^1_{1/2} = t^2_{1/2}$. We assume that the sample sizes in both groups are equal to N, that all subjects have the same number of repeated measurements obtained at the same time intervals, and that the between-subject variance of the true rate parameter is equal to σ_s^2 for

both populations. It is shown in Appendix I that for fixed values of k and D, the required number of subjects per group for attaining a $(1-\beta)$ level of power to detect the alternative hypothesis, $H_A : t^1_{1/2} \neq t^2_{1/2}$ at an α significance level is

$$N = \ln(2)^2 \left(\frac{z_{\alpha/2}\sqrt{2/\bar{\Lambda}^4} + z_\beta \sqrt{\frac{1}{\lambda_1^4} + \frac{1}{\lambda_2^4}}}{(t^1_{1/2} - t^2_{1/2})} \right)^2 \left[\sigma_s^2 + \frac{\sigma_e^2 12(k-1)}{D^2 k(k+1)} \right], \tag{9}$$

where $z_{\alpha/2}$ and z_β denote the standard normal deviates corresponding to $\alpha/2$ and β significance levels, respectively, and $\bar{\Lambda} = (\lambda_1 + \lambda_2)/2$.

Note that since the required sample size depends on λ_1 and λ_2, the actual values of $t^1_{1/2}$ and $t^2_{1/2}$ need to be specified, and not just the magnitude of their difference. Equation (9) can also be easily re-expressed to determine the k or D to attain a specified level of power, for fixed values of the other parameters.

The formula for determining the sample size was derived assuming that the duration of the study and the number of repeats are fixed. However, the most common situation when designing a study is that k and D, in addition to N, need to be determined. Methods similar to the above may be utilized to calculate the optimal values for the number of subjects, number of repeats, and duration of study which will minimize the overall study cost, while attaining a specified level of power. The total cost of the study can be denoted as $C = c_0 + (c_1 k + c_2 D + c_3)2N$, where c_0 denotes overhead and other fixed costs which are independent of k, D, and N; c_1 and c_2 are the costs associated with each measurement and each interval of follow-up, respectively; and c_3 denotes the cost of enrolling each additional subject.

The optimal parameter values for k, D, and N can be determined by minimizing C, subject to the constraint in (9). Unlike the previous problem, however, this has no closed form solution and must be solved iteratively. A Newton-Raphson algorithm, written in SAS PROC IML, was utilized to estimate the optimal parameters (Press et al., 1986). This algorithm requires calculation of the first and second order derivatives, with respect to the

parameters of interest, of the function which is to be minimized. In this case, the function is $C = c_0 + (c_1 k + c_2 D + c_3)2N$, with N substituted by the expression in (9). Expressions for the first and second-order derivatives of C with respect to k and D, and details of the algorithm are given in Appendix II.

Specific values of $z_\alpha, z_\beta, \sigma_s^2, \sigma_e^2, t_{1/2}^1$, and $t_{1/2}^2$, as well as the costs, c_1, c_2, and c_3, must be assumed. Note that because the first and second-order derivatives of C with respect to k and D are independent of c_0, the overhead cost will not affect the outcome of the minimization process, and hence, need not be specified. Given initial starting values for k and D, the algorithm iteratively finds the values which minimize C. The optimal number of subjects, N, is then calculated from (9). An example illustrating the methods is presented in the next section.

3 Example

In this section, utilization of the methods to design a study to compare the differences in the half-life of PCBs between subjects with and without breast cancer will be illustrated. First, values of the variance components, σ_s^2, the between-subject variance in the true rate parameter, and σ_e^2, the variance of the deviations of the observed measurements (log transformed) from the values predicted from equation (2), must be assumed. Variance estimates were obtained using pilot data from the NYU Women's Health Study (NYUWHS), a prospective cohort of 14,291 women who have been donating multiple blood samples over time (Toniolo et al, 1991). A breast cancer case-control study nested in this cohort found elevated, but non-significant, levels of PCBs measured at enrollment among cases relative to controls (Wolff et al, 1993). No half-lives were measured at that time because only one blood donation per subject was analyzed. Subsequently, pilot data became available on subjects in the NYUWHS who had at least 3 blood donations. Concentrations of PCBs were measured in serum specimens that have been collected and stored since enrollment; the assays were

performed under the direction of Dr. Mary Wolff at Mt. Sinai Medical Center. Details of the experimental protocol are provided in Wolff et al (1991).

In calculating the half-lives for this cohort, the concentrations of PCBs within subjects are assumed to be decreasing over time. In principle, however, the body burden of PCBs may actually increase in individuals who are chronically exposed to low levels of the toxin and whose initial concentrations were in the range of normal background levels, resulting in negative half-life estimates. For our example, the analysis was restricted to include only the 15 subjects with at least 3 measurements of PCBs available who had a positive estimate of half-life. The mean half-life of PCBs among these subjects was estimated to be 10 years.

An estimate of σ_e^2 was obtained by fitting the following linear mixed ANOVA model:

$$Y_{ij} = \mu + \alpha_i + \lambda_i t_{ij} + e_{ij}, \tag{11}$$

where Y_{ij} is defined as the logarithm of the j^{th} measurement of PCB from subject i, μ denotes the overall mean, α_i denotes a random subject effect, λ_i is the rate parameter for subject i, t_{ij} is the time since enrollment for subject i and donation j, and e_{ij} is the residual error, which is assumed to be distributed with mean 0, and common variance, σ_e^2. The mean squared error resulting from model (11) estimates σ_e^2. Fitting (11) to the NYUWHS data yielded $\hat{\sigma}_e^2 = .046$.

Obtaining an estimate of the between-subject variance of the true rate parameters, σ_s^2 was more problematic. If the measurements from all subjects were made at the same set of time points, $\mathbf{t} = \{t_1, ..., t_k\}$, then one could estimate σ_s^2 by first estimating λ_i for all subjects and subtracting $\frac{\hat{\sigma}_e^2}{\sum_{j=1}^{k}(t_j - \bar{t})^2}$ from the observed variance of $\hat{\lambda}_i$, since the unconditional variance of $\hat{\lambda}_i$ is equal to $\sigma_s^2 + \frac{\hat{\sigma}_e^2}{\sum_{j=1}^{k}(t_j - \bar{t})^2}$. However, in the NYUWHS and in most other studies, subjects have different numbers of repeated measurements obtained at varying time intervals. In this case, a conservative estimate of σ_s^2 would be to use the observed variance of $\hat{\lambda}_i$. Although this leads to an overestimate of the required sample size, the approximation improves as the number of repeated measurements and the duration between measurements become large.

The observed variance of the rate parameters of PCBs from our pilot data was estimated to be .0028.

Before determining the optimal design for comparing the half-lives between two populations, we illustrate how one can generate tables using (8) and the estimates of σ_s^2 and σ_e^2 to evaluate the effect of increasing k, D and N on the precision of the estimate of a single population half-life. Suppose one assumes that the true underlying half-life of PCB for the breast cancer cases is 11 years. This corresponds to a population rate parameter of $\lambda_p = \ln(2)/11 = .063$. Using (8), we generated Table 2, which shows the variance of $\hat{t}_{1/2}^P$ for selected values of k, D and N. For example, with a sample size of 75 subjects measured four times over a period of 8 years, the variance of the estimated half-life will be 1.66, corresponding to a 95% confidence interval width of: $2 \times 1.96 \times \sqrt{1.66} = 5.05$ years for the true population half-life. In this particular example, increasing the duration of study by a given number of years, say x, results in greater gains in precision compared with increasing by x the number of repeats or number of subjects. This result, however, may not apply for different values of σ_s^2 and σ_e^2.

The optimal design for comparing the population half-lives of PCB between breast cancer cases and controls will now be determined. The following values for the costs of the study were assumed: \$200 for each PCB assay (c_1), \$25 for each year of follow-up (c_2), and \$75 to enroll each subject (c_3). Assuming that the half-life of PCB among control subjects is 8 years and that the study should have 80% power to detect an increase in the half-life to 11 years among breast cancer cases at an $\alpha = .05$ significance level, we found, using the iterative algorithm described in Appendix II, that the optimal design is to enroll 100 subjects per group, and to obtain 2 measurements per subject over 12 years.

Even though this design is the one which will minimize the overall cost of the study, in practice, it may not be feasible to conduct the study over a time period as long as 12 years. Suppose that 5 years is the maximum feasible duration of study. Then, one can minimize C with respect to k and N, while keeping D fixed at 5 years, to obtain the optimal design

for a 5 year study. Iterative methods similar to the above were used to determine that the optimal design for a 5 year study is to obtain 2 measurements per subject on 186 subjects per group. Although this design will yield the same level of power over a shorter duration as the first design, it will cost an additional \$6275.

Figure 1 shows how the optimal k, D and N change as a function of the cost of the assay, assuming the values of the other parameters have not changed. For example, if the cost of the PCB assay were only \$2 rather than \$200, then the optimal design is to obtain 26 measurements per subject over 5 years and enroll 103 subjects per group. The greatest changes in the optimal values for k, D and N occur when $c1$ ranges from \$1-\$9. For assay costs greater than \$9, the optimal value for k remains stable at 2 measurements. Corresponding changes in the optimal D and N in this region of $c1$ are minimal. Similar graphs can be generated to evaluate the impact of varying the values of the other parameters on the optimal values.

It is straightforward to show that specification of the level of power, type I error rate, and population half-lives only influence the determination of the optimal N, and not k and D (see Appendix II). Thus, in order to evaluate how the optimal design changes as a function of α, $1-\beta$, $t^1_{1/2}$, and $t^2_{1/2}$, one need only to re-calculate N using (9), since the required k and D will remain unchanged. For instance, continuing the initial example from above, in order for the study to attain 70%, as opposed to 80% power, the required number of subjects is reduced to 77 per group, while the optimal k and D remain as above ($k = 2; D = 12$). The values for k and D are affected only by the costs, c_1, c_2 and c_3, and the values of the variance components, σ^2_s and σ^2_e.

4 Conclusions

Understanding the pharmacokinetics, and in particular, the rate of excretion from the body of environmental contaminants is crucial for ascertaining the etiologic role of these risk factors

in the development of disease. In this paper, methods for designing studies on estimating and comparing the half-lives of environmental toxins have been described. The ability to utilize these methods, however, may be limited by the availability of preliminary estimates for the variance components. Although most studies on population half-lives provide estimates of the variance of the population rate parameters, which may be used as an upper bound estimate of σ_s^2, estimates of σ_e^2 are rarely published. The availability of pilot data becomes especially important in this case. Also, because iterative methods are required to determine the optimal design for comparing two population half-lives, the techniques may not be easily implemented in practice for some investigators and a statisician may need to be consulted. Finally, the techniques in this paper are based on the assumptions of a one-compartment exponential decay model and a linear least-squares regression estimate of the rate parameter, λ. Thus, they cannot be applied to the multi-compartment case. Extension of this work to accommodate the multi-compartment assumption will be the subject of future research.

Most published reports on the half-lives of environmental contaminants have been based on small numbers of subjects and small numbers of repeated measurements. The large variability in the published estimates of the half-lives of toxins such as PCB may reflect the lack of precision that results from inadequate study designs. This paper demonstrates the gains in precision and statistical power that may be achieved by increasing the sample size, number of repeats, and time interval between repeats, and underscores the importance of study design when planning studies on half-life.

Table 1: Values of ω ($\times 10$) as a function of the number of repeats, k, and the duration of study, D.

	D									
k	1	2	3	4	5	6	7	8	9	10
2	9.609	2.402	1.068	0.601	0.384	0.267	0.196	0.150	0.119	0.096
3	9.609	2.402	1.068	0.601	0.384	0.267	0.196	0.150	0.119	0.096
4	8.648	2.162	0.961	0.541	0.346	0.240	0.176	0.135	0.107	0.086
5	7.687	1.922	0.854	0.480	0.307	0.214	0.157	0.120	0.095	0.077
6	6.864	1.716	0.763	0.429	0.275	0.191	0.140	0.107	0.085	0.069
7	6.177	1.544	0.686	0.386	0.247	0.172	0.126	0.097	0.076	0.062
8	5.605	1.401	0.623	0.350	0.224	0.156	0.114	0.088	0.069	0.056
9	5.125	1.281	0.569	0.320	0.205	0.142	0.105	0.080	0.063	0.051
10	4.717	1.179	0.524	0.295	0.189	0.131	0.096	0.074	0.058	0.047
15	3.363	0.841	0.374	0.210	0.135	0.093	0.069	0.053	0.042	0.034
20	2.608	0.652	0.290	0.163	0.104	0.072	0.053	0.041	0.032	0.026

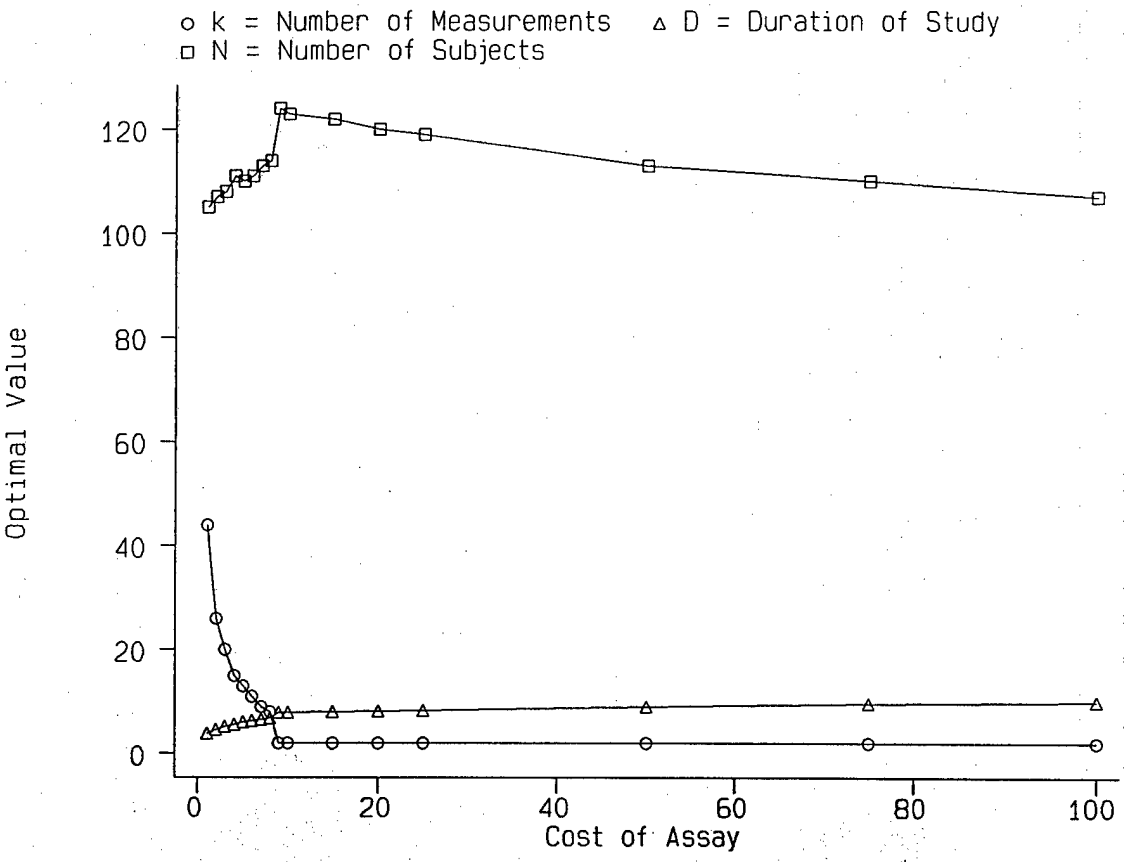
Table 2: Values of $V(\hat{t}^P_{1/2})$ for $N = 25, 50, 75, 100$; $\sigma^2_s = .0028$; $\sigma^2_e = .046$; $\lambda_p = .063$, as a function of the number of repeats, k, and duration of study, D.

N = 25							
	D						
k	2	4	6	8	10	15	20
2	31.48	10.43	6.53	5.17	4.54	3.91	3.70
4	28.67	9.73	6.22	4.99	4.43	3.86	3.67
6	23.46	8.43	5.64	4.67	4.22	3.77	3.62
8	19.78	7.51	5.23	4.44	4.07	3.71	3.58
10	17.19	6.86	4.95	4.28	3.97	3.66	3.55
15	13.24	5.87	4.51	4.03	3.81	3.59	3.51
20	11.03	5.32	4.26	3.89	3.72	3.55	3.49

N = 50							
	D						
k	2	4	6	8	10	15	20
2	15.74	5.22	3.27	2.58	2.27	1.96	1.85
4	14.33	4.86	3.11	2.50	2.21	1.93	1.83
6	11.73	4.21	2.82	2.33	2.11	1.89	1.81
8	9.89	3.75	2.62	2.22	2.04	1.85	1.79
10	8.60	3.43	2.47	2.14	1.98	1.83	1.78
15	6.62	2.94	2.25	2.01	1.90	1.80	1.76
20	5.52	2.66	2.13	1.95	1.86	1.78	1.75

N = 75							
	D						
k	2	4	6	8	10	15	20
2	10.49	3.48	2.18	1.72	1.51	1.30	1.23
4	9.56	3.24	2.07	1.66	1.48	1.29	1.22
6	7.82	2.81	1.88	1.56	1.41	1.26	1.21
8	6.59	2.50	1.74	1.48	1.36	1.24	1.19
10	5.73	2.29	1.65	1.43	1.32	1.22	1.18
15	4.41	1.96	1.50	1.34	1.27	1.20	1.17
20	3.68	1.77	1.42	1.30	1.24	1.18	1.16

N = 100							
	D						
k	2	4	6	8	10	15	20
2	7.87	2.61	1.63	1.29	1.13	0.98	0.92
4	7.17	2.43	1.56	1.25	1.11	0.97	0.92
6	5.86	2.11	1.41	1.17	1.05	0.94	0.90
8	4.95	1.88	1.31	1.11	1.02	0.93	0.89
10	4.30	1.71	1.24	1.07	0.99	0.92	0.89
15	3.31	1.47	1.13	1.01	0.95	0.90	0.88
20	2.76	1.33	1.07	0.97	0.93	0.89	0.87


o k = Number of Measurements
△ D = Duration of Study
□ N = Number of Subjects
Optimal Value
120
100
80
60
40
20
0
0
20
40
60
80
100
Cost of Assay

References

Chen, P.H., Luo, M.L., Wong, C.K., Chen, C.J. (1982) Comparative rates of elimination of some individual polychlorinated biphenyls from the blood of PCB-poisoned patients in Taiwan. Fd Chem Toxicol 20:417-425.

Cox, D.R. and Hinkley, D.V. (1974) Theoretical Statistics. New York: Chapman and Hall.

Elo O., Vuojolahti, P., Janhunen, H., Rantanen, J. (1985) Recent PCB accidents in Finland. Environ Health Persp 60:315-319.

Lawton, R.W., Brown, J.F. Jr., Ross, M.R., Feingold, J. (1985) Comparability and precision of serum PCB measurements. Arch Environ Health 40:29-37.

Phillips, D.L., Smith, A.B., Burse, V.W., Steele, G.K., Needham, L.L., Hannon, W.H. (1989) Half-life of polychlorinated biphenyls in occupationally exposed workers. Arch Environ Health 44:351-354.

Phillips, D.L. (1989) Propagation of error and bias in half-life estimates based on two measurements. Arch Environ Contam Toxicol 18:508-514.

Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T. Numerical Recipes. Cambridge University Press, 1986, Cambridge.

Schlesselman, J.J. (1973) Planning a longitudinal study: frequency of measurement and study duration. J Chron Dis 26:561-570.

Steele, G., Stehr-Green, P., Welty, E. (1986) Estimates of the biologic half-life of polychlorinated biphenyls in human serum. New Engl J Med 314:926-927.

Toniolo, P., Pasternack, B., Shore, R., Sonnenschein, E., Koenig, K., Rosenberg, C., Strax, P., Strax, S. Endogenous hormones and breast cancer: a prospective cohort study. Breast Cancer Res Treat 1991; 18:S23-S26.

Wolff, M.S., Rivera, M., Baker, D.B. (1991) Detection limits of organochlorine pesticides and related compounds in blood serum. Bull Environ Contam Toxicol 47:499-503.

Wolff, M.S., Toniolo, P.G., Lee, E.W., and Dubin, N. (1993) Blood levels of organochlorine residues and risk of breast cacner. J Natl Cancer Inst 85:648-652.

Yakushiji, T., Watanabe I., Kuwabara, K., Tanaka, R., Kashimoto, T., Kunita, N., Hara, I. (1984) Rate of decrease and half-life of polychlorinated biphenyls in the blood of mothers and their children occupationally exposed to PCBs. Arch Environ Contam Toxicol 13:341-345.

Appendix I: Determination of the Sample Size for Comparing Two Population Half-lives

We assume that the sample sizes from the two populations are the same and are equal to N. Let $\hat{t}^1_{1/2}$ and $\hat{t}^2_{1/2}$ denote the observed half-lives in the two populations, and $\hat{\lambda}_1$ and $\hat{\lambda}_2$ denote the estimates of the corresponding rate parameters. Then the test statistic for evaluating $H_0 : t^1_{1/2} = t^2_{1/2}$ is of the form:

$$Z = \frac{\hat{t}^1_{1/2} - \hat{t}^2_{1/2}}{\sqrt{\frac{1}{N}\left[\sigma_s^2 + \frac{\sigma_e^2 12(k-1)}{D^2 k(k+1)}\right] \ln(2)^2 (2/\bar{\lambda}^4)}},$$

where Z is distributed as $N(0,1)$, and $\bar{\lambda} = (\hat{\lambda}_1 + \hat{\lambda}_2)/2$.

If the test statistic is to have power $(1-\beta)$ to detect the alternative hypothesis, $H_A : t^1_{1/2} > t^2_{1/2}$ at a 1-sided $\alpha = .05$, then we have the following expression:

$$Pr\left\{ \frac{\hat{t}^1_{1/2} - \hat{t}^2_{1/2}}{\sqrt{\frac{1}{N}\left[\sigma_s^2 + \frac{\sigma_e^2 12(k-1)}{D^2 k(k+1)}\right] \ln(2)^2 (2/\bar{\lambda}^4)}} > z_\alpha \middle| H_A \right\} = 1-\beta, \qquad (1)$$

where z_α denotes the critical value corresponding to the α proportion in the upper tail of the standard normal distribution.

After some algebra, (1) can be re-expressed as:

$$\begin{aligned} 1-\beta \;=\; & Pr\left\{ \frac{\hat{t}^1_{1/2} - \hat{t}^2_{1/2} - (t^1_{1/2} - t^2_{1/2})}{\sqrt{\frac{1}{N}\left[\sigma_s^2 + \frac{\sigma_e^2 12(k-1)}{D^2 k(k+1)}\right] \ln(2)^2 (\frac{1}{\lambda_1^4} + \frac{1}{\lambda_2^4})}} \right. \\ & \left. > \frac{z_\alpha \sqrt{\frac{1}{N}\left[\sigma_s^2 + \frac{\sigma_e^2 12(k-1)}{D^2 k(k+1)}\right] \ln(2)^2 (2/\bar{\lambda}^4)} - (t^1_{1/2} - t^2_{1/2})}{\sqrt{\frac{1}{N}\left[\sigma_s^2 + \frac{\sigma_e^2 12(k-1)}{D^2 k(k+1)}\right] \ln(2)^2 (\frac{1}{\lambda_1^4} + \frac{1}{\lambda_2^4})}} \right\} \end{aligned}$$

Under the alternative hypothesis, the expression on the left-hand side of the inequality has a $N(0,1)$ distribution. Thus,

$$1-\beta = Pr\left\{ Z > \frac{z_\alpha \sqrt{\frac{1}{N}\left[\sigma_s^2 + \frac{\sigma_e^2 12(k-1)}{D^2 k(k+1)}\right] \ln(2)^2 (2/\bar{\lambda}^4)} - (t^1_{1/2} - t^2_{1/2})}{\sqrt{\frac{1}{N}\left[\sigma_s^2 + \frac{\sigma_e^2 12(k-1)}{D^2 k(k+1)}\right] \ln(2)^2 (\frac{1}{\lambda_1^4} + \frac{1}{\lambda_2^4})}} \right\}.$$

Note that the definition of $\bar{\lambda}$ requires knowledge of $\hat{\lambda}_1$ and $\hat{\lambda}_2$, which are available only after completion of the study. However, for large n, $\bar{\lambda}$ may be well approximated by $\bar{\Lambda} = (\lambda_1 + \lambda_2)/2$. After substituting $\bar{\Lambda}$ for $\bar{\lambda}$ above, setting the expression on the right-hand side equal to $-z_\beta$ and solving for N, we have

$$N = \frac{\ln(2)^2 \left[z_\alpha \sqrt{2/\bar{\Lambda}^4} + z_\beta \sqrt{\frac{1}{\lambda_1^4} + \frac{1}{\lambda_2^4}} \right]^2}{(t^1_{1/2} - t^2_{1/2})} \left[\sigma_s^2 + \frac{\sigma_e^2 12(k-1)}{D^2 k(k+1)} \right]. \qquad (2)$$

This sample size was derived under the assumption of a one-sided alternative hypothesis. When H_A is two-sided, the required sample size is obtained by simply substituting $z_{\alpha/2}$ for z_α in (2).

Appendix II: Details of the Newton-Raphson Algorithm

The overall cost of the study is equal to :

$$\begin{aligned} C &= c_0 + (c_1 k + c_2 D + c_3) 2N \\ &= c_0 + (c_1 k + c_2 D + c_3) 2 \left\{ \frac{\ln(2)^2 \left[z_\alpha \sqrt{2/\bar{\Lambda}^4} + z_\beta \sqrt{\frac{1}{\lambda_1^4} + \frac{1}{\lambda_2^4}} \right]^2}{(t_{1/2}^1 - t_{1/2}^2)} \right\} \left[\sigma_s^2 + \frac{\sigma_e^2 12(k-1)}{D^2 k(k+1)} \right], \end{aligned} \tag{2}$$

with N substituted by the expression in (10). The optimal k and D which will minimize C are the values which will solve the following first derivative equations:

$$\frac{\partial C}{\partial k} = A \left\{ c_1 \sigma_s^2 + \frac{c_1 \sigma_e^2 12}{D^2} \left[\frac{2}{(k+1)^2} \right] + \left[\frac{c_3 \sigma_e^2 12}{D^2} + \frac{c_2 \sigma_e^2 12}{D} \right] \left[\frac{1 + 2k - k^2}{(k^2+k)^2} \right] \right\} = 0$$

$$\frac{\partial C}{\partial D} = A \left\{ c_2 \sigma_s^2 - \frac{12 \sigma_e^2 (k-1)}{k(k+1)} \left(\frac{2c_1 k}{D^3} + \frac{c_2}{D^2} + \frac{2c_3}{D^3} \right) \right\} = 0,$$

where $A = 2 \frac{\ln(2)^2 \left[z_\alpha \sqrt{2/\bar{\Lambda}^4} + z_\beta \sqrt{1/\lambda_1^4 + 1/\lambda_2^4} \right]^2}{(t_{1/2}^1 - t_{1/2}^2)}$. Note that since A is not a function of k and D, the constant can be omitted without affecting the final solution.

The Newton-Raphson method for solving the above equations requires calculation of the corresponding second-order derivatives:

$$\frac{\partial^2 C}{\partial k^2} = \frac{12 \sigma_e^2}{D^2} \left[\frac{-4c_1}{(k+1)^3} + \frac{(c_3 + c_2 D) 2 (k^3 - 3k^2 - 3k - 1)}{(k^2+k)^3} \right]$$

$$\frac{\partial^2 C}{\partial D^2} = \frac{24(k-1)\sigma_e^2}{(k+1)k} \left(\frac{3c_1 k}{D^4} + \frac{c_2}{D^3} + \frac{3c_3}{D^4} \right)$$

$$\frac{\partial^2 C}{\partial k \partial D} = -12 \sigma_e^2 \left[\frac{4c_1}{(k+1)^2 D^3} + \frac{1 + 2k - k^2}{(k^2+k)^2} \left(\frac{2c_3}{D^3} c_2 D^2 \right) \right].$$

Given the preliminary values, (k_0, D_0), the algorithm calculates updated values for k and D according to:

$$\begin{bmatrix} k \\ D \end{bmatrix} = \begin{bmatrix} k_0 \\ D_0 \end{bmatrix} - \begin{bmatrix} \frac{\partial^2 C}{\partial k^2} & \frac{\partial^2 C}{\partial k \partial D} \\ \frac{\partial^2 C}{\partial D \partial k} & \frac{\partial^2 C}{\partial D^2} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial C}{\partial k} \\ \frac{\partial C}{\partial D} \end{bmatrix}$$

The algorithm repeatedly updates (k, D) and calculates the above until convergence is obtained.

Chapter IV

The Analysis of Correlated Panel Data Using a Continuous Time Markov Model

1 Introduction

Many biological processes may be described in terms of a finite number of states which individuals proceed through over time. For example, Klein et al (1984) and Kay (1986) have modeled the phases of cancer as a three stage disease process.

In most longitudinal studies which follow the passage of individuals through various biological states, the available data consist of the series of states each subject was observed to be in at various points in time. Because monitoring is not continuous, information about the process is usually unavailable between follow-up times, and the exact times of transition from one state to another are not known. This type of data, in which the observations consist of the states occupied by the individuals under study at a sequence of discrete time points, is often referred to as "panel" data. Methods for analyzing panel data using time-homogeneous Markov models have been explored by Klein et al. (1984), Kalbfleisch and Lawless (1985), and Gentleman et al (1994).

Existing methods for the analysis of panel data have been applicable to the case where a single biological process is of interest. However, many clinical studies monitor two or more processes over time on each individual, where the processes observed within a subject may be correlated. For example, in certain longitudinal studies involving HIV-infected subjects, several immunologic variables, such as CD4, CD8 and serum immunoglobulins are each measured periodically on all subjects. Also, in most clinical trials, several possibly correlated clinical variables are repeatedly assessed during the follow-up period of each patient.

In the clinical trial which motivated the methodology proposed in this paper, patients with the Acquired Immune Deficiency Syndrome (AIDS) were randomized to receive one of two drugs: ganciclovir or foscarnet, for the treatment of cytomegalovirus (CMV) retinitis, and were periodically monitored for the development of toxicities in several hematologic, enzymatic, and chemical variables. For each variable, five states of toxicity were defined: none, mild, moderate, severe and life-threatening, according to the AIDS Clinical Trials Group (ACTG) toxicity grading system. The data for a

particular patient consisted of repeated assessments over time of the toxicity states of each of several possibly correlated clinical variables, yielding what we refer to as "correlated panel" data. To our knowledge, statistical methods for analyzing correlated panel data have not been previously developed.

Wei, Lin and Weissfeld (1989) described a method for analyzing multivariate survival data which models each survival distribution marginally using a Cox proportional hazards model, and makes no specific assumptions regarding the structure of dependence between distinct failure times on each subject. This approach has also been utilized by Stram, Wei and Ware (1988). In this paper, we apply the marginal approach to the analysis of correlated panel data. We model each process based on a time-homogeneous Markov model allowing for covariates and impose no specific dependence structure among the related processes. The resulting estimators are shown to be asymptotically jointly normal with a covariance matrix that accounts for the dependence among related processes and can be consistently estimated. Simultaneous inference procedures are also proposed. In Section 2, the method for analyzing correlated panel is described, and in Section 3 an example is presented using data from the CMV clinical trial.

2 Methods

2.1 Markov Modeling and Estimation

Assume that K processes are periodically observed on each subject, and that each process has Q ordered states. We model the k^{th} process, $(k = 1, ..., K)$, as a time-homogeneous Markov process, where the first $Q-1$ are transient states, and the Qth is a single absorbing state. Transitions are assumed to occur irreversibly from one state to the next. The time-homogeneous Markov assumption implies that times to transition between states are exponentially distributed with a hazard, or transition intensity, that may be modeled as a function of covariates.

For notational convenience, we assume a single covariate, z, is of interest. Then, following the proportional hazards model of Cox (1972), the transition intensities for each process can be modeled as,

$$\lambda_q^k(z) = \lambda_q^k \exp(\beta_q^k z), \qquad q = 1, ..., Q-1, k = 1, ..., K$$

where λ_q^k is the baseline transition intensity between states q and $q+1$ for the kth process, and β_q^k is the regression parameter denoting the effect of z on λ_q^k.

For the i^{th} subject ($i = 1, ..., n$) with covariate z_i, the probability that the k^{th} process will be in state s at time t_1+t, given that the process was in state r at time t_1, is denoted as $p_{rs}^k(t, z_i)$. Then, under the assumptions that $\lambda_j^k(z_i) \neq \lambda_l^k(z_i)$, $l \neq j, j, l = 1, ..., Q-1$, standard methods may be utilized to show that this transition probability is equal to

$$p_{rs}^k(t, z_i) = (-1)^{s-r} \lambda_r^k(z_i) ... \lambda_{s-1}^k(z_i) \sum_{j=r}^{s} \left[\frac{e^{-\lambda_j^k(z_i)t}}{\prod_{l=r, l\neq j}^{s} \{\lambda_j^k(z_i) - \lambda_l^k(z_i)\}} \right], \qquad (1)$$

where $r = 1, ..., Q-1; s \geq r$, and $s < Q$ (Chiang, 1980). The transition probabilities from a transient state r to the absorbing state Q are equal to

$$p_{rQ}^k(t, z_i) = 1 - \sum_{j=r}^{Q-1} p_{rj}^k(t, z_i), \; r = 1, ...Q-1,$$

and $p_{QQ}^k(t, z_i) = 1$.

Now, let s_{ij}^k denote the state of the kth process for the ith subject at the jth observation ($j = 1, ., ., m_i$), at time t_{ij}^k, measured in days since date of initial observation. Then the observed history of the k^{th} process for the i^{th} subject may be denoted as

$$A_i^k = (s_{i1}^k, t_{i1}^k, ..., s_{im_i}^k, t_{im_i}^k).$$

Following the Markov property and conditional probability arguments, the probability of observing A_i^k, conditional on z_i, can be expressed as,

$$\begin{aligned} \Pr(A_i^k | z_i) &= \Pr(k^{th} \text{ process in state } s_{i1}^k \text{ at time } t_{i1}^k, ..., \text{state } s_{im_i}^k \text{ at time } t_{im_i}^k | z_i) \\ &= \prod_{j=1}^{m_i - 1} p_{s_{ij}, s_{i,j+1}}^k (t_{i,j+1}^k - t_{ij}^k, z_i), \end{aligned}$$

where the form of $p^k_{s_{ij},s_{i,j+1}}(t^k_{i,j+1}-t^k_{ij},z_i)$ may be derived from (1). Then the marginal likelihood function for the k^{th} process over all n subjects is

$$\begin{aligned} L(\theta_k) &= \prod_{i=1}^{n} Pr(A_i^k|z_i) \\ &= \prod_{i=1}^{n}\left\{\prod_{j=1}^{m_i-1} p^k_{s_{ij},s_{i,j+1}}(t^k_{i,j+1}-t^k_{ij},z_i)\right\}, \end{aligned}$$

where $\theta_k=(\lambda_1^k,...\lambda_{Q-1}^k\beta_1^k,...\beta_{Q-1}^k)$. The corresponding log-likelihood is equal to

$$\mathcal{L}(\theta_k)=\sum_{i=1}^{n}\left[\sum_{j=1}^{m_i-1}\log\{p^k_{s_{ij},s_{ij+1}}(t^k_{ij+1}-t^k_{ij},z_i)\}\right]. \tag{2}$$

Substituting (1) in (2), the maximum likelihood estimator, $\hat{\theta}_k$, of θ_k, for each process may be obtained by solving $\partial\mathcal{L}(\theta_k)/\partial\theta_k=0$, for $k=1,...,K$. It follows that under the usual regularity conditions, $\sqrt{n}(\hat{\theta}_k-\theta_k)$ for each process is asymptotically normal with mean 0 and covariance matrix $I_k^{-1}(\theta_k)$, where $I_k(\theta_k)$ is defined as

$$I_k(\theta_k)=E_{\theta_k}\left[-\frac{1}{n}\frac{\partial^2\mathcal{L}(\theta_k)}{\partial\theta_k^2}\right]=E_{\theta_k}\left[\frac{1}{n}\left\{\frac{\partial\mathcal{L}(\theta_k)}{\partial\theta_k}\right\}\left\{\frac{\partial\mathcal{L}(\theta_k)}{\partial\theta_k}\right\}^T\right]. \tag{3}$$

The estimates of θ_k from different processes may be correlated. To take into account the correlations among $\hat{\theta}_k$'s, consider the Taylor series expansion

$$\sqrt{n}(\hat{\theta}_k-\theta_k)\sim\mathcal{I}_k^{-1}(\theta_k)\frac{U_k}{\sqrt{n}},$$

where

$$\frac{1}{\sqrt{n}}U_k=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial\mathcal{L}_i(\theta_k)}{\partial\theta_k}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_{ik}, \tag{4}$$

and

$$\mathcal{I}_k(\theta_k)=-\frac{1}{n}\frac{\partial^2\mathcal{L}(\theta_k)}{\partial\theta_k^2}.$$

By the above development, $\sqrt{n}\left\{\begin{pmatrix}\hat{\theta}_1\\ \vdots\\ \hat{\theta}_K\end{pmatrix}-\begin{pmatrix}\theta_1\\ \vdots\\ \theta_K\end{pmatrix}\right\}$ can be written as the sum of independent vectors with mean 0, and thus, under the usual regularity conditions,

follow a multivariate normal distribution with mean 0 and covariance matrix Σ, which may be partitioned as

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1K} \\ \Sigma_{12} & \Sigma_{22} & \cdots & \Sigma_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{1K} & \Sigma_{2K} & \cdots & \Sigma_{KK} \end{bmatrix},$$

where $\Sigma_{kk} = I_k^{-1}(\theta_k)$, and Σ_{jk} is the covariance matrix between $\sqrt{n}(\hat{\theta}_j - \theta_j)$ and $\sqrt{n}(\hat{\theta}_k - \theta_k)$. From (4), the covariance matrix of $(U_k/\sqrt{n},\ U_j/\sqrt{n})$ can be estimated consistently by

$$\frac{1}{n}\sum_{i=1}^{n} \hat{\xi}_{ik}\hat{\xi}_{ij}^T, \tag{5}$$

where $\hat{\xi}_{ik}$ is equal to $\partial \mathcal{L}_i(\theta_k)/\partial\theta_k$ evaluated at $\hat{\theta}_k$, and $\mathcal{L}_i(\theta_k)$ is the contribution to the log-likelihood made by the i^{th} subject. The term $\hat{\xi}_{ij}$ is defined similarly. Since $\mathcal{I}_k^{-1}(\hat{\theta}_k)$ and $\hat{I}_k^{-1}(\hat{\theta}_k) = \left(\frac{1}{n}\sum_{i=1}^{n} \hat{\xi}_{ik}\hat{\xi}_{ik}^T\right)^{-1}$ are equivalent in probability, it follows that one can estimate Σ_{jk} by

$$\frac{1}{n}\hat{I}_j^{-1}(\theta_j)\left(\sum_{i=1}^{n} \hat{\xi}_{ij}\hat{\xi}_{ik}^T\right)\hat{I}_k^{-1}(\theta_k),$$

for $j \neq k$ and $j, k = 1, ..., K$.

2.2 Hypothesis Testing

The multivariate normal distribution of $(\hat{\theta}_1, ..., \hat{\theta}_K)$ provides a basis for making simultaneous inferences about the θ_ks. In particular, suppose that we are interested in evaluating the effect of z_i on the transition intensities for the K processes. Let $\eta = \{\eta_1, ..., \eta_{K\times(Q-1)}\} = (\beta_{11}, ..., \beta_{Q-1,1}, ..., \beta_{1K}, ..., \beta_{Q-1,K})$ denote the parameters of interest, and let $\hat{\Phi}$ denote the estimate of the covariance matrix of $\hat{\eta}$, which can be obtained from $\hat{\Sigma}$. Then asymptotically the following holds

$$W = \hat{\eta}'\hat{\Phi}^{-1}\hat{\eta} \sim \chi^2_{\{K\times(Q-1)\}}. \tag{6}$$

The above W statistic can then be used to simultaneously test the null hypotheses H_l: $\eta_l = 0$, for $l = 1, ..., K \times (Q-1)$. If the null hypotheses are rejected, the next step is to decide which of the η_l are non-zero.

Let $\{\tilde{\eta}_1, \tilde{\eta}_2, ..., \tilde{\eta}_{K\times(Q-1)}\}$ denote the standardized estimator of $\{\hat{\eta}_1, \hat{\eta}_2, ..., \hat{\eta}_{K\times(Q-1)}\}$, where $\tilde{\eta}_l = \hat{\eta}_l/\hat{\phi}_{ll}^{1/2}$, and $\hat{\phi}_{ll}$ is the $(l,l)^{th}$ element of $\hat{\Phi}$. Under the null hypotheses H_l, $\{\tilde{\eta}_1, ..., \tilde{\eta}_{K\times(Q-1)}\}$ is approximately normal with mean 0 and covariance matrix $\tilde{\Phi} = \{\tilde{\phi}_{pq}\}$, where $\tilde{\phi}_{pq} = \hat{\phi}_{pq}/(\hat{\phi}_{pp}\hat{\phi}_{qq})^{1/2}$, and $\hat{\phi}_{pq}$ is the (p,q)th element in $\hat{\Phi}$. A conventional multiple testing procedure rejects $H_l, l = 1, ..., K \times Q - 1$, if $|\tilde{\eta}_l| > c$, where c is the smallest number such that

$$Pr(|\tilde{\eta}_l| < c, l = 1, ..., K \times Q | H_1, ..., H_{K\times Q-1}) \leq 1 - \alpha,$$

and α is a prespecified level of signifcance. However, the sequential multiple test procedures studied by Marcus, Peritz and Gabriel (1976), Holm (1979) and Wei and Stram (1988) will yield more powerful tests than the above conventional multiple testing procedures.

Following Wei and Stram (1988), let $\tilde{\eta}_l^*$ be the lth largest absolute value of the $\tilde{\eta}_l$'s, and let $\tilde{\Phi}^*$ be the corresponding variance-covariance matrix that is obtained by rearranging the rows and columns of $\tilde{\Phi}$. Also, let $H_l^* : \eta_l^* = 0$ be the ordered hypotheses from the H_l's according to the order of $\tilde{\eta}_1^*, ..., \tilde{\eta}_{K\times Q-1}^*$. Furthermore, let $(Z_1, ..., Z_{K\times Q-1})$ denote a multivariate normal vector with mean zero and covariance matrix $\tilde{\Phi}^*$. Starting with the hypothesis H_1^*, we reject $H_l^*, l = 1, ..., K \times Q - 1$, if $P(\min_{l\leq j\leq K\times Q-1} Z_j \leq -|\tilde{\eta}_l^*|) \leq \alpha/2$, where α is a prespecified two-sided level of significance, provided that $H_1^*, ..., H_{l-1}^*$ have been tested and rejected. It can be shown that the asymptotic type I error probability of this procedure is α for any combination of true H_ls. An illustration of the sequential multiple testing procedure is provided in the example.

3 Example

The methods developed in the previous section were applied to data derived from the CMV retinitis clinical trial. This trial, conducted by the Studies of Ocular Complications of AIDS Research Group, was designed to compare ganciclovir (Cytovene, Syntex Laboratories) and foscarnet (Foscavir, Astra Pharmaceutical Products) in the

treatment of CMV retinitis in AIDS patients (SOCA, 1992).

Two hundred and thirty-four patients were randomly assigned to treatment: 127 to ganciclovir and 107 to foscarnet. One of the objectives of this trial was to compare the toxicity profiles of the two drugs. We focus particularly on the toxic effects of the treatments on the two hematologic variables: hemoglobin and absolute neutrophil count, which could be correlated within individuals. Five stages of toxicity were defined for each outcome based on the following ACTG toxicity criteria. For HGB, the stages were defined as, none: $\geq 11.0g/dl$, mild: $9.5 - 10.9g/dl$, moderate: $8.0 - 9.4g/dl$, severe: $6.5 - 7.9g/dl$ and life-threatening:$< 6.5g/dl$. The corresponding toxicity grades for ANC were, none: $\geq 1500/\mu l$, mild: 1000-1499/μl, moderate: 750-999/μl, severe: $500 - 749/\mu l$, and life-threatening: $< 500/\mu l$. It should be emphasized that even though the quantitative levels of each variable were known, the grade of toxicity rather than the actual level was more clinically relevant. In addition, because very few subjects were observed to progress to the life-threatening stage of both HGB and ANC toxicity, this state was combined with the severe toxicity state to yield four possible states for each process. The goals of the analysis were to estimate and compare the effects of foscarnet and ganciclovir on the transition intensities between the states of toxicity of HGB and ANC, and to estimate the average waiting time in each state.

The methods in the previous section were based on the assumption that transitions between states occur irreversibly. Levels of HGB and ANC may not decrease monotonically over time, but could increase from one visit to the next. When this occurred in our data set, the process was simply held in the state defined by the previous observation until it was observed to decline again, thus allowing us to model the states as irreversible. The frequencies of occurrences in which the processes were observed to increase rather than decrease from one visit to the next were relatively low and were similar in the two groups (ANC: 20% vs. 18% for foscarnet and ganciclovir, respectively; HGB: 21% vs. 18% for foscarnet and ganciclovir, respectively), so that the assumption of the progressive model would not likely bias the estimates of treatment effects.

The covariate denoting treatment group for the i^{th} subject is defined as $z_i = 0$ for Foscarnet, and $z_i = 1$ for Ganciclovir. The parameter vector for HGB is denoted as $\theta_1 = (\lambda_1^1, \lambda_2^1, \lambda_3^1, \beta_1^1, \beta_2^1, \beta_3^1)$, and the corresponding vector for ANC is $\theta_2 = (\lambda_1^2, \lambda_2^2, \lambda_3^2, \beta_1^2, \beta_2^2, \beta_3^2)$, where β_q^k is the coefficient corresponding to the effect of treatment on the transition intensity between states q and $q+1$ for process k.

Tables 1 and 2 show the estimates of the transition intensities and the β coefficients for the two processes. The mean waiting time for the k^{th} state was estimated by $1/(\hat{\lambda}e^{z\hat{\beta}})$. The mean waiting times in each state of ANC for ganciclovir are less than half the mean waiting times for foscarnet, indicating the toxic effects on ANC are more severe with ganciclovir. On the other hand, for HGB, foscarnet appears to be associated with a higher probability of transition from mild to moderate, and moderate to severe toxicity.

Figures 1 and 2 show the estimated cumulative distribution functions of time to severe toxicity of HGB and ANC, respectively, for each treatment. The difference between treatments in the probability of reaching severe toxicity is much greater for ANC than HGB. The Kaplan-Meier estimates of the distribution functions are also shown for comparison. Our model appears to underestimate the probabilities of progressing to severe toxicity relative to the Kaplan-Meier estimates. The lack-of-fit of the model was also apparent in results (not shown) of the goodness-of-fit test proposed by Gentleman (1994) that was performed. The inadequacy of the assumed model could indicate that the estimates of the regression coefficients and corresponding covariance matrix may not be valid; however, for illustrative purposes we proceed to present the results of the sequential multiple testing procedure.

The covariance matrix for $(\hat{\beta}_1^1, \hat{\beta}_2^1, \hat{\beta}_3^1, \hat{\beta}_1^2, \hat{\beta}_2^2, \hat{\beta}_3^2)$ was estimated to be

$$\begin{pmatrix} 0.0560 & 0.0020 & 0.0023 & 0.0006 & 0.0116 & -0.0009 \\ & 0.0689 & -0.0006 & 0.0087 & 0.0072 & -0.0004 \\ & & 0.1072 & 0.0088 & -0.0023 & 0.0082 \\ & & & 0.0861 & 0.0024 & -0.0014 \\ & & & & 0.0692 & 0.0082 \\ & & & & & 0.0968 \end{pmatrix}.$$

To jointly test $H_l^k : \beta_l^k = 0, l = 1, 2, 3; k = 1, 2$, we calculated the test statistic, W, from (4), which approximately follows a χ^2 distribution with 6 degrees of freedom. The observed W was 42.96 which is significant at less than 1% level.

In order to evaluate which of the coefficients were non-zero, we employed the multiple testing procedure described in the previous section. The standardized parameter estimates $(\tilde{\beta}_1^1, \tilde{\beta}_2^1, \tilde{\beta}_3^1, \tilde{\beta}_1^2, \tilde{\beta}_2^2, \tilde{\beta}_3^2)$ were 3.477, 3.910, 2.629, -1.721, -1.039, and 0.426 respectively. Since $-|\tilde{\beta}_2^1| < -|\tilde{\beta}_1^1| < -|\tilde{\beta}_3^1| < -|\tilde{\beta}_1^2| < -|\tilde{\beta}_2^2| < -|\tilde{\beta}_3^2|$, we tested the null hypotheses in the order: $H_2^1, H_1^1, H_3^1, H_1^2, H_2^2, H_3^2$. Using the integration algorithm for multivariate normal probabilities in Schervish(1984, 1985), we have

$$\begin{aligned} P(\min\{Z_1, Z_2, Z_3, Z_4, Z_5, Z_6\} \leq -3.910) &\approx 0.0002 \\ P(\min\{Z_2, Z_3, Z_4, Z_5, Z_6\} \leq -3.477) &\approx 0.0012 \\ P(\min\{Z_3, Z_4, Z_5, Z_6\} \leq -2.629) &\approx 0.0279 \\ P(\min\{Z_4, Z_5, Z_6\} \leq -1.721) &\approx 0.1230. \end{aligned}$$

Assuming a two-sided $\alpha = 0.05$, these results suggest that ganciclovir has a greater toxic effect on ANC than foscarnet, across all states of toxicity. However, the evidence is insufficient to conclude that the toxicity profiles for HGB are different for the two treatments.

4 Conclusions

Although the proposed methods were applied to data from a clinical trial of two treatments for CMV retinitis, they are also applicable to longitudinal studies of other diseases, such as breast cancer, in which several clinical variables are periodically monitored over time on each subject.

A key assumption of the methods is that transition rates between states do not change over time. This time-homogeneous Markov assumption is quite restrictive, however, and may in part explain the apparent lack of fit of the model. An alternative approach is to assume that the intensities are piecewise constant over distinct time intervals (Gentleman et al., 1994). Kalbfleisch and Lawless (1985) also suggested fitting a parametric time-dependent model.

In addition, we have assumed a progressive Markov model, in which transitions between states can only occur in one direction. The proposed methodology can in principle be extended to the situation where arbitrary transitions are allowed, by modelling each process marginally using the methods in Kalbfleisch and Lawless (1985) and deriving the joint covariance matrix estimate of all model parameters empirically. However, because a larger number of parameters are required to specify the reversible model compared with the progressive model, estimation of the parameters may be more difficult.

ACKNOWLEDGEMENTS

This research was supported by Grants AI-27742 and DA-04722 from the National Institutes of Health, and DAMD17-94-J-4174 from the Department of Defense. We are grateful to the Studies of the Ocular Complications of AIDS Research Group (SOCA) for providing the data in the example. The analyses and interpretations herein are solely those of the authors and not of the SOCA research group as a whole, and

representatives of the SOCA Research Group have not reviewed or checked the analyses or interpretations so presented for clarity or accuracy or for compliance with the stated principles of SOCA.

REFERENCES

Chiang, C.L. (1980) *An Introduction to Stochastic Processes and Their Applications,* 2nd edn, Krieger, New York.

Cox, D.R.(1972), "Regression Models and Life-Tables"(with discussion), *J urnal of the Royal Statistical Society,* Ser. B. 34, 187-220.

Gentleman, R.C., Lawless, J.F., Lindsey, J.C. and Yan, P.(1994), "Multi-State Markov Models For Analysing Incomplete Disease History Data With Illustraions For HIV Disease," *Statistics in Medicine*, 13, 805-821.

Holm, S.(1979), "A Simple Sequentially Rejective Multiple Test Procedure," *Scandinavian Journal of Statistics*, 6, 65-70.

Kalbfleisch, J.D. and Lawless, J.F.(1985), "The Analysis of Panel Data Under A Markov Assumption," *Journal of the American Statistical Association*, 80, 863-871.

Kay, R.(1986), "A Markov Model for Analysing Cancer Markers and Disease States in Survival Studies," *Biometrics*, 42, 855-865.

Klein, J.P., Klotz, J.H. and Grever, M.R.(1984), "A Biological Marker Model for Predicting Disease Transitions," *Biometrics*, 40, 927-936.

Longini Jr. I.M., Clark, W.S., Byers, R.H., Ward, J.W., Darrow, W.W., Lemp, G.F. and Hethcote, H.W.(1989), "Statistical Analysis of the Stages of HIV Infection Using a Markov Model," *Statistics in Medicine*, 8, 831-843.

Marcus, R., Peritz, E. and Gabriel, K.R.(1976), "On Closed Testing Procedures with Special Reference to Ordered Analysis of Variance," *Biometrika*, 63, 655-660.

Schervish, M.J. (1984), "Multivariate Normal Probabilities with Error Bound, "*Applied Statistics*, 33, 81-94.

Stram, D.O., Wei, L.J., and Ware, J.H.(1988), "Analysis of Repeated Ordered Categorical Outcomes with Possibly Missing Observations and Time-Dependent Covariates," *Journal of the American Statistical Association*, 83, 631-637.

Studies of Ocular Complications of AIDS Research Group (1992), "Mortality in Patients With the Acquired Immunodeficiency Syndrome Treated With Either Foscarnet or Ganciclovir For Cytomegalovirus Retinitis," *The New E gland d urnal of Medicine*, 326:213-220.

Wei, L.J., Lin, D.Y. and Weissfeld, L.(1989), "Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distribution," *d urnal of the American Statistical Association*, 84, 1065-1073.

Wei, L.J. and Stram, D.O.(1988), "Analysing Repeated Measurements with Possibly Missing Observations by Modeling Marginal Distributions," *Statistics in Medicine*, 7, 139-148.

Table-1. Estimated parameters and mean waiting times in each state of toxicity for Absolute neutrophil count

Stage	($\lambda \pm$S.D.,$\beta \pm$S.D.)	Mean waiting time for Fos.(days)	Mean waiting time for Gov.(days)
1	(.0070±.0013, .823±.237)	(143 ± 27)	(63 ± 8)
2	(.0059±.0014, 1.03±.263)	(169 ± 40)	(61 ± 8)
3	(.0065±.0020, .861±.327)	(154 ± 47)	(65 ± 8)

Table-2. Estimated parameters and mean waiting times in each state of toxicity for Hemoglobin count

Stage	($\lambda \pm$S.D.,$\beta \pm$S.D.)	Mean waiting time for Fos.(days)	Mean waiting time for Gov.(days)
1	(.0139±.0027, -.505±.293)	(72 ± 14)	(119 ± 26)
2	(.0098±.0018, -.273±.263)	(102 ± 19)	(135 ± 25)
3	(.0062±.0013, 0.133±.311)	(161 ± 34)	(142 ± 32)

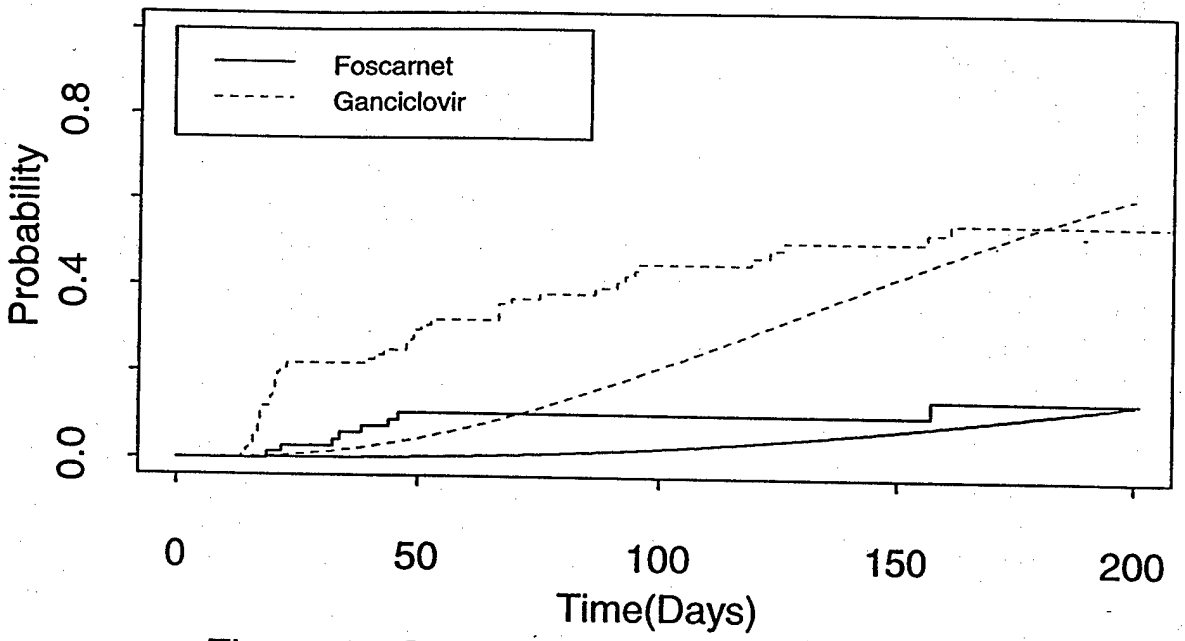


Figure 1. C.D.F. and Kaplan-Meier Estimates of Time to Severe Toxicity of ANC

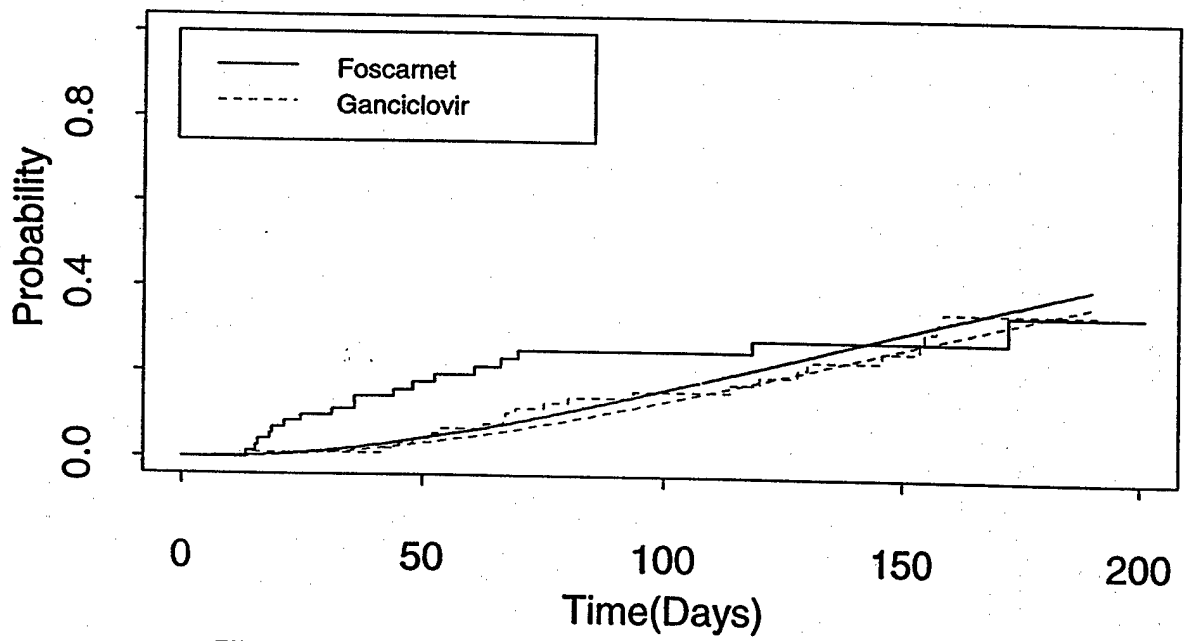


Figure 2. C.D.F. and Kaplan-Meier Estimates of Time to Severe Toxicity of HGB

Chapter V

The Application of Equivalence Tests in Validation Studies of Dietary Assessment Methods

1 Introduction

The role of diet in the development of diseases such as cancer and heart disease is the focus of numerous epidemiologic studies. The ability of these studies to evaluate the true relationship between diet and disease depends on having valid information on an individual's habitual dietary intake. Obtaining accurate data on dietary intake can be challenging, however, and efforts continue to be directed toward developing and improving instruments which are not only informative, but also cost-effective and easy to administer.

The main approach for evaluating a new method for assessing dietary intake is to perform a validation study. Given that a subject's "true" dietary intake can never be known, a validation study involves comparing the new technique with an established superior method. For example, food frequency questionnaires, which have the advantage of being inexpensive and easy to administer, are often validated against the more accurate but also more costly and time-consuming dietary record (Pietinen et al., 1988, Rimm et al, 1992). If the observed degree of agreement between the two methods is "sufficient", then investigators conclude that the new instrument is suitable for use in epidemiologic studies of nutrition and disease.

The most appropriate statistical technique for assessing agreement between two dietary methods has been debated extensively in the literature. Many investigators utilize the correlation coefficient to measure agreement, whereas others have advocated comparing the means. In addition, the relevance of hypothesis tests in validation studies has been questioned, and *a priori* criteria for judging validity are either inappropriate or absent. For example, when means are compared, the typical conclusion is that the two methods are comparable if the null hypothesis of no difference is not rejected– n improper statistical procedure. When correlations coefficients are calculated, rejection of the null hypothesis that $\rho = 0$ is irrelevant, and there is no statistically-based inferential criterion to determine when the correlation is high enough to be acceptable as a demonstration of validity. Hence, conclusions from validity studies have been either fuzzy or inappropriate.

In this paper, we present a conceptual approach for drawing inferences regarding the validity of a dietary assessment method based on equivalence testing. This approach was originally developed for the analysis of therapeutic equivalence trials (Blackwelder, 1982). The objective of a therapeutic equivalence trial is to show that a new treatment is as effective as a standard, in contrast to the goal of most conventional clinical trials which is to prove superiority of the experimental treatment. The demonstration of therapeutic equivalence between treatments is of interest when the new regimen offers advantages such as reduced toxicity, invasiveness, and cost. The goal of most validation studies of dietary instruments can be viewed analogously: to prove that a technique that is simpler, easier to administer, and more cost effective is as informative as an established method of assessing dietary intake.

Because of the parallels between the objectives of the two type of studies, statistical methods for proving the equivalence between two therapeutic regimens are also applicable to proving the validity of a new dietary assessment method. In the following sections, we describe examples where conventional hypothesis tests have been misused in validation studies, and then present an alternative approach based on an equivalence testing framework. In this approach, one must define in advance the degree of agreement which constitutes adequate validity. Appropriately specified hypothesis tests or confidence intervals are then used to infer whether this magnitude of agreement has been demonstrated. Implications for determining the sample size of a validation study based on the proposed method are also discussed.

Use of Conventional Hypothesis Tests in Validation Studies

The most common approach for evaluating the degree of agreement between two dietary methods is to calculate the correlation coefficient, which quantifies the degree of linear association between measurements obtained by the two methods. Although this statistic cannot detect the presence of a systematic bias, the correlation does provide a measure of how similarly the two methods position their measurements relative to their corresponding means (Kelsey, Thompson, and Evans, 1986).

Many investigators only report estimates of the correlations and do not perform statistical tests because quantifying the degree of measurement error is viewed as the main objective in a validation study rather than hypothesis testing (Willett et al, 1987; Rimm et al, 1992). Others have reported p-values for the null hypothesis, $H_o : \rho = 0$, and p-values which are less than 0.05 have been interpreted as indicating appreciable agreement between dietary methods (refs). As Garrow (1995) and Margetts and Thompson (1995) have pointed out, however, statistical significance does not necessarily imply the existence of a correlation of sufficient magnitude; only that the correlation is not likely to be 0. For example, an observed correlation as low as 0.1 would be statistically significant (p=0.046) if based on a validation sample size of 400 subjects. Therefore, tests of the conventional hypothesis of a null correlation are not very meaningful in validation studies.

The use of the correlation coefficient as a measure of agreement has been criticized because of the lack of generalizability of this parameter to other study populations with different magnitudes of nutritional variability and the inability of the correlation to capture a systematic bias. An alternative method for assessing the agreement between two methods is to compare the means. Statistical tests have also been misused with this approach, however. For example, Chu et al (1984) compared the mean dietary intakes from two methods and if they "did not differ significantly ($p > 0.05$) ... then the two methods agreed absolutely". This is not a valid approach for evaluating agreement because the p-value depends on the sample size and the lack of statistical significance could be due to an insufficient number of subjects rather than true equivalence in the two methods.

Lee et al (1983) proposed the use of regression coefficients for analyzing data from validation studies and claim that two methods are in complete agreement only if they show a linear regression coefficient that is not statistically different from 1.0. But similar to the previous example, one could erroneously conclude agreement between the methods simply by having a small validation study which has inadequate power to detect a non-unity slope (Wahrendorf, 1985).

It is clear that conventional tests of significance are not appropriate for assessing validity, since rejecting the hypothesis of a zero correlation does not necessarily imply sufficient agreement, nor does failing to detect a difference in means or a non-unity slope. On the other hand, reporting only the estimated level of agreement is not satisfactory either, since any evaluation of agreement should factor in not only the magnitude of the estimate, but also the associated degree of imprecision. Statistical methods developed by Dunnett and Gent(1977) and Blackwelder(1982) for evaluating the equivalence of two treatments in a clinical trial are an alternative approach for evaluating agreement between two dietary instruments in a validation study.

2 Methods

Equivalence Testing Framework

In conventional hypothesis testing involving two groups, the goal is to demonstrate that a difference between the groups exists. Thus, the null hypothesis is specified in terms of equivalence and the alternative is set up in the direction of a difference, e.g., $H_0 : \mu_1 = \mu_2$ and $H_A : \mu_1 \neq \mu_s$, where μ_1 and μ_2 denote the means in two groups. In contrast, when one is trying to demonstrate equivalence between groups, this formulation of the null and alternative hypotheses no longer applies since the p-value for the above will be "a measure of the evidence against the null hypothesis not for it, and insufficient evidence to reject the null hypothesis does not imply sufficient evidence to accept it." (Blackwelder,1982).

To overcome this difficulty, the strategy in equivalence studies is to reverse the traditional roles of the null and alternative hypotheses so that the null hypothesis now specifies non-equivalence between the two groups, whereas the alternative is in the direction of equivalence. To see how this principle can be extended to a validation study, suppose the objective of the study is to show that the means of two dietary methods are equivalent or nearly equivalent.

One first specifies δ, the maximum allowable difference between the methods that would still make them "equivalent" for practical purposes. Then, one formulates the null hypothesis so that the absolute difference in means is greater than or equal to δ, i.e., $H_0^* : |\mu_1 - \mu_2| \geq \delta$ and the alternative so that $H_A^* : -\delta < \mu_1 - \mu_2 < \delta$. Rejecting H_0^* then allows one to conclude that the magnitude of the difference in means is within a tolerable range, and therefore that the methods can be considered interchangeable.

Because measurements obtained from the experimental and established methods in validation studies are paired, H_0^* can be evaluated by performing a paired t-test of a non-null difference using the following statistic:

$$t = \frac{|\bar{d}| - \delta}{s_d/\sqrt{n}},$$

where $\bar{d}$ denotes the mean difference between the methods, s_d denotes the standard deviation of the difference, and n is the sample size. The two-sided p-value is equal to $2 \times \Pr(T > |t|)$, where T follows a t distribution with $(n-1)$ degrees of freedom.

As an illustration, Willett et al (1987) validated a semi-quantitative food frequency questionnaire against a 1 year dietary record using a group of 27 subjects. No hypothesis testing was performed and only means, standard deviations, and correlations were reported for various nutrients. For vitamin C, the population mean based on the dietary record was estimated to be 125 mg/day with a standard deviation of 87 mg/day, whereas the food frequency questionnaire yielded a mean of 146 mg/day with standard deviation of 88 mg/day. Since the standard deviation of the difference was not reported in the paper, we approximated it by the following formula: $\sqrt{s_1^2 + s_2^2 - 2\rho s_1 s_2}$, and used the available estimates of the individual standard deviations and correlation given in the paper. The correlation was reported to be 0.38, yielding an estimate of the standard deviation of the difference of 97 mg/day. We performed a test of the the conventional null hypothesis of no difference in means, and the two-sided p-value is 0.27. Based on the non-significance of this result, an investigator may erroneously infer that the two instruments are equivalent in measuring vitamin C.

Now suppose that a maximum difference of 30 mg/day in vitamin C between the two methods is considered tolerable, which corresponds to half of the recommended daily allowance(RDA) of this nutrient. Then the two-sided p-value corresponding to the null hypothesis: $H_0^* : |\mu_1 - \mu_2| \geq 30$ versus the alternative: $H_A^* : -30 < \mu_1 - \mu_2 < 30$ is equal to 0.64. This result indicates that under the assumed criterion for equivalence in vitamin C, the evidence is inadequate to conclude that the magnitude of the difference between the two methods is sufficiently small to be acceptable,. On the other hand, if a less stringent criterion for equivalence were used, such as $\delta = 60$ mg/day, then the corresponding p-value for the null hypothesis, $H_0^* : |\mu_1 - \mu_2| \geq 60$ is equal to 0.047. The null hypothesis for this definition of δ would be rejected, and one would conclude that the two dietary instruments were equivalent for measuring vitamin C.

The principles of equivalence testing can also be extended to the case where a correlation is used to measure the strength of agreement in a validation study. In the previous example, a smaller difference in means corresponded to increased validity. In contrast, a higher degree of correlation implies greater agreement, so δ here should be defined as the minimum acceptable level of correlation. One possible value for δ is 0.4, since "levels of validity lower than this will rather seriously attenuate associations" (Willett, 1990). However, higher values should be used if one wants to be more conservative.

Since two methods would be judged to be non-equivalent if $\rho \leq \delta$ and equivalent for practical purposes if $\rho > \delta$, the null and alternative hypotheses should be specified as, $H_0^* : \rho \leq \delta$ and $H_A^* : \rho > \delta$, respectively. Note that the direction of this null hypothesis based on a correlation is opposite the one based on the absolute difference in means. Again, this is because the null hypothesis in an equivalence study should always be specified in the direction of non-equivalence.

Using Fisher's arctanh transformation, the hypothesis of a non-null correlation may be

tested using the following statistic which follows a standard normal distribution:

$$z = [C(r) - C(\delta)]\sqrt{(N-3)},$$

where $C(x) = 0.5\log_e(\frac{1+x}{1-x})$, and r denotes the observed correlation. Because H_0^* is one-sided for a correlation, the corresponding p-value is equal to $1 - \Phi(z)$, where $\Phi()$ denotes the cumulative standard normal distribution. Rejecting H_0^* allows one to conclude that there is an acceptable magnitude of agreement between the two methods.

To illustrate this approach, suppose a validation study based on 25 subjects was performed and the observed correlation coefficient was 0.6. Given the magnitude of this point estimate, and that the two-sided p-value for the classical null hypothesis of no correlation, $H_0 : \rho = 0$ is highly statistically significant (p=0.0012), an investigator may conclude that the experimental instrument is valid. However, the p-value associated with the hypothesis, $H_0^* : \rho \leq 0.4$ is 0.10. According to this result, one cannot claim that validity has been demonstrated because the hypothesis that the correlation is below the minimum acceptable level cannot be rejected.

Defining δ will be the most difficult and subjective aspect of the equivalence testing approach for assessing validity. This is especially so when the comparison of means is used to evaluate agreement since a different δ needs to be specified for each nutrient and relevant information on which to base the definitions may be lacking. Another potential difficulty is that standard statistical software packages do not generally provide p-values for tests of non-null differences in means and non-null correlations. Thus, the investigator may have to compute the relevant test statistic by hand or with the aid of a calculator. When a large number of equivalence tests need to be performed for various nutrients, it may be more efficient to create a simple computer program to perform the calculations.

Confidence Interval Approach to Assessing Validity

Confidence intervals convey all the information of a hypothesis test, but are additionally

useful because they provide a range of values of the true level of agreement that are consistent with the observed data. Durrleman and Simon (1990) state that "such approach is appealing because it emphasizes the magnitude of the difference...rather than its statistical significance, which is often misunderstood or misused as a binary decision rule". For this reason, many investigators favor the use of confidence intervals over hypothesis testing to evaluate equivalence.

Equivalence between two methods based on confidence intervals is assessed by determining whether every point within the range of the confidence interval for the true level of agreement is acceptable. If so, then the two methods may be considered equivalent.

Continuing the example from the previous section involving a correlation, the lower bound of the one-sided 95% confidence interval for the true correlation is 0.33 based on an observed estimate of 0.6 from 25 subjects. Since 0 is below the lower confidence bound, one can conclude that some level of association between the methods is probable. However, given that 0.4 is the minimum acceptable correlation, validity has not been demonstrated because correlations as low as 0.33 cannot be ruled out based on these data.

Similarly, the two-sided 95% confidence interval for the true difference in means of vitamin C measured by the food frequency questionnaire and the dietary record studied by Willett is (-17.4, 59.4). Agreement between the methods can be inferred only if all values within this range correspond to a difference of no practical importance.

The use of confidence intervals in validation studies has also been advocated by Burema, van Staveren and Feunekes (1995). Bland and Altman (1986) proposed a similar approach based on calculating "limits of agreement" for the difference in means. They advocated that one compute $\bar{d} \pm 1.96 s_d$, where $\bar{d}$ denotes the mean difference in the two methods, and s_d is the standard deviation of the difference. If all individual differences within the range of these limits of agreement are acceptable, the authors suggest that the two methods can be used interchangeably.

Although the general idea of this technique is similar to the confidence interval approach

to evaluating equivalence, the boundaries on the limits of agreement and the confidence interval can differ quite markedly. Since the confidence interval for the true difference based on the normal approximation: $\bar{d} \pm 1.96 s_d/\sqrt{n}$, is computed using the standard error of the mean difference rather than the standard deviation, Bland and Altman's limits of agreement will be much wider than the confidence interval, and therefore more conservative for assessing validity, especially with large sample sizes. In particular, in the vitamin C example, the limits of agreement are (-169.1, 211.1), which is substantially wider than the 95% confidence interval for the difference computed above. A bias as large as 211 mg/day in Vitamin C intake, which is greater than three times the RDA, would have to be acceptable in order to conclude equivalence between the methods based on the Bland and Altman method.

Bland and Altman's limits of agreement apply primarily to differences between the two instruments for individuals rather than mean differences. This approach is therefore applicable to evaluate a new method's use in making individual clinical assessments. However, when trying to assess in a validation study whether a systematic bias exists between the methods, where bias is defined as a difference in the means of the two methods, the confidence interval approach is more appropriate since it conveys the range of plausible values for the true difference in means.

Sample Size Estimation

Suppose one wanted to design a validation study in which two means would be compared to assess agreement. The goal is to ensure that the study will have adequate power to conclude that the two methods are interchangeable within a maximum tolerable difference, δ, under the assumption that the true difference between methods is equal to Δ. In equivalence studies, power is generally evaluated at the point of exact equivalence, i.e., $\Delta = 0$. The required sample size in order for the study to have $(1 - \beta)\%$ power at a two-sided α level

can be approximated using the following formula:

$$N = \frac{\sigma_d^2(z_{\alpha/2} + z_\beta)^2}{(\delta - \Delta)^2}, \tag{1}$$

where σ_d^2 is the variance of the difference in measurements obtained by the two methods, and $z_{\alpha/2}$ and z_β denote the $\alpha/2$ and β quantiles of the standard normal distribution, respectively. If an estimate of σ_d^2 is not available from previous studies, it can be approximated by assuming the variances of the measurements from the two methods are equal to a common value, σ, and using the formula,

$$\sigma_d^2 = 2\sigma^2(1 - \rho),$$

with estimates of σ^2 and ρ.

It is easily shown that formula (1) applies whether the hypothesis testing or confidence interval approaches to evaluating validity are assumed when designing the study. Furthermore, when $\Delta = 0$, expression (1) equals the sample size formula for paired observations used in conventional superiority studies. The interpretation of δ differs, of course, in the two studies. Whereas in equivalence studies, δ is defined as the maximum tolerable difference between groups, in superiority studies it is the difference that is considered clinically or scientifically important to detect. Usually, δ is smaller when trying to prove equivalence rather than superiority, since investigators are generally reluctant to tolerate large differences between groups. Because the sample size is inversely proportional to the square of δ, this implies that the required sample sizes for equivalence studies tend to be larger than superiority studies for the same level of power and Type I error rate.

To determine the sample size of a validation study utilizing a correlation as the measure of agreement, assume that we expect the true correlation to be ρ_A. Then, in order for the study to have $(1-\beta)\%$ power to conclude that the true correlation is greater than a minimum δ at a one sided α level, the required number of subjects based on Fisher's transformation is:

$$N = \left(\frac{z_\alpha + z_\beta}{C(\delta) - C(\rho_A)}\right)^2 + 3.$$

For example, if the true correlation is expected to be 0.6 and the minimum acceptable correlation is 0.4, then one would need approximately 90 subjects for the study to have 80% power to conclude validity at a one-sided $\alpha = 0.05$ level. Under similar assumptions, but using a different approach, Willett (1990) recommended sample sizes of about 100-200 subjects for validation studies.

3 Conclusions

In this paper, we have presented statistical methods for inferring whether the validity of a new dietary-intake instrument has been proven in a validation study. Our approach is also applicable to nearly all kinds of validity studies employed in epidemiology, not just diet studies, e.g. to questions on smoking, hormone use, sunlight exposure, or surrogate vs. self respondents.

In planning the analyses of a validity study, a necessary step is to consider what kind of equivalence is important. Does the absolute level of the two measures need to be equivalent, or is it only important that the rank order of subjects on the two measures be very similar? These questions will determine what equivalence tests are planned.

The proposed methods were illustrated using two common statistics for assessing validity: the correlation coefficient and the difference in means. However, the methods can also be readily extended to other statistics for measuring agreement between dietary instruments, provided the distribution of the statistic is known or can be approximated. For example, one can easily apply the principles of equivalence testing to the kappa statistic, which measures the degree of agreement between categories of nutritional intake, since it is known to have a large sample normal distribution with asymptotic variance given in Fleiss, Cohen and Everitt (1969).

This approach can also be applied to reproducibility studies, in which one defines ad-

equate reproducibility as an intraclass correlation greater than some defined value. The formulas for calculating the intraclass correlation, corresponding variance, and confidence interval needed to apply the methods are given in Donner (1986).

Either of the proposed approaches for assessing validity, based on hypothesis tests or confidence intervals, is an improvement over current practices in the analysis of validation studies which include computing estimates of agreement without information on the degree of imprecision, or performing inappropriate hypothesis tests. The confidence interval approach may be preferable to significance testing for assessing validity because the former is more informative and easier to interpret. With the significance testing approach, the meaning of the p-value depends on the definition of δ. This may be confusing to readers since p-values are classically associated with conventional hypothesis tests rather than tests of non-null values. To minimize this confusion, investigators need to be very explicit in defining the acceptable level of agreement when reporting the p-value of an equivalence test.

Acknowledgments

This work was supported by Grants numbers DAMD17-94-J-4174 from the Department of Defense Breast Cancer Research Program, CA... from the National Cancer Institute, ...

References

Blackwelder, W.C. (1982) Proving the null hypothesis in clinical trials. Controlled clinical trials 3:345-353.

Bland, J.M. and Altman, D.J. (1986) Statistical methods for assessing the agreement between two methods of clinical measurement. Lancet 1:307-310.

Burema, J., van Staveren, W.A., Feunekes, G.I.J. (1995) Guidelines for reports on validation studies. European Journal of Clinical Nutrition 49:932-933.

Chu, S.Y., Kolonel, L.N., Hankin, J.H., and Lee, J. (1984) A comparison of frequency and quantitative dietary methods for epidemiologic studies of diet and disease. American Journal of Epidemiology 119;323-34.

Dunnett, C.W. and Gent, M. (1977) Significance testing to establish equivalence between treatments, with special reference to data in the form of 2 x 2 tables. Biometrics 33:593-602.

Donner, A. (1986) A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. International Statistical Review 54:67-82.

Durrleman and Simon (1990)Planning and Monitoring of Equivalence Studies. Biometrics 46:329-336.

Fleiss, J.L., Cohen, J., and Everitt, B.S. (1969) Large-sample standard errors of kappa and weighted kappa. Psychol. Bull., 72:232-327.

Garrow, J.S.(1995)Validation of methods for estimating habitual diet: proposed guidelines. European Journal of Clinical Nutrition 49:231-232.

Kelsey, J.L., Thompson, W.D., Evans, A.S. (1986) Methods in Observational Epidemiology. New York: Oxford University Press.

Lee, J. Kolonel, L.N., and Hankin, J.H. (1983) On establishing the interchangeablity of different dietary-intake assessment methods used in studies of diets and cancer. Nutritional Cancer 5:215-218.

Margetts, B.M. and Thompson, R.L. (1995) Validation of dietary intake estimation (letter). European Journal of Clinical Nutrition 49:934.

Pietinen, P., Hartman, A.M., Haapa, E., Rasanen, L., Haapakoski, J., Palmgren, J., Albanes, D., Virtamo, J., and Huttunen, J.K. (1988) Reproducibility and validity of dietary assessment instruments. American Journal of Epidemiology 128:544-555.

Rimm, E.B., Giovannucci, E.L., Stampfer, M.J., Colditz, G.A., Litin, L.B. and Willett, W.C. (1992) Reproducibility and validity of an expanded self-administered semi-quantitative food frequency questionnaire among male health professionals. American Journal of Epidemiology 135:1114-1126.

Wahrendorf, J. (1985) Re: A comparison of frequency and quantitative methods for epidemiologic studies of diet and disease (letter). American Journal of Epidemiology 121:776.

Ware, J.H. and Antman, E.M. (1997) Equivalence Trials. New England Journal of Medicine 337:1159-1161

Willett, W., Reynolds, R.D., Cottrell-Hoehner, S., Sampson, L., and Browne, M.L. (1987) Validation of a semi-quantitative food frequency questionnaire: comparison with a 1 year diet record. Journal of the American Dietetic Association 87:43-47.

Willett, W. Nutritional Epidemiology (1990) Oxford University Press, New York, N.Y.

Chapter VI

Adjusting Hormone Levels for Day of Menstrual Cycle in Studies of Breast Cancer and Hormones in Pre-Menopausal Women

1 Introduction

Although levels of prolactin and bioavailable estradiol appear to be relatively stable over the phases of a woman's menstrual cycle, other hormones, such as total estradiol, fluctuate considerably. (Toniolo et al., 1993; Koenig et al., 1993.; Wu et al., 1976; Takatani et al., 1991). Epidemiologic studies investigating the association of total estradiol and risk of breast cancer among premenopausal women must adjust a subject's hormone level for day of cycle either in the design or analysis stage of the study, in order for the comparisons between cases and controls to be valid.

In the NYU Women's Health Study, a nested case-control study of serum hormonal levels and breast cancer, one of the criteria for matching controls with a breast cancer case among pre-menopausal women was the day of menstrual cycle on which the first blood specimen was collected, measured in number of days prior to next expected onset of menses. Subsequent blood donations, however, could not be matched on day of cycle. Therefore, a method was needed to standardize hormone measurements obtained at different times during the menstrual cycle for subjects in the same matched set.

Rosenberg et al (1994) used the first measurement from each control subject to fit a three-piece spline model to describe the change in total estradiol level over the menstrual cycle. For each subject, the estradiol measurement adjusted for day of cycle was then calculated as the difference between the observed value and the expected value from this calibration curve, measured in units of standard deviation. The limitation with this approach, however, is that because only the first measurement from each subject was used to fit the calibration curve, the curve is estimated with less precision than one that is estimated using all available repeated measurements. In addition, the width of the confidence intervals for the relative risks for breast cancer based on the adjusted estradiol measurements are underestimated, since they do not take into account the extra variation due to estimation of the parameters

of the calibration curve.

We propose an alternative method for describing the within-subject change in estradiol levels over the menstrual cycle, based on a mixed linear model with cubic splines, which utilizes all the repeated measurement data for each subject. The use of cubic splines in the model yields a smoother curve than the one fit by Rosenberg et al, which was based on a three-piece spline: two parabolas and a straight line, without smoothed join points.

We use the results from the mixed linear model to adjust each subject's hormone level for day-of-cycle. The adjusted measurement then becomes the exposure in a conditional logistic regression analysis. Bootstrap methods are utilized to obtain estimates of the corresponding 95% confidence intervals for the regression coefficients which account for the variation in the estimated calibration curve.

2 Methods

Let $\mathbf{y}_i = \{y_{i1}, ..., y_{ik_i}\}$ denote the vector of hormone levels for the i^{th} woman measured on k_i occasions for $i = 1, ...n$. Furthermore, let $\mathbf{t}_i = \{t_{i1}, ..., t_{ik_i}\}$ denote the vector of the number of days prior to next menses at which the $\mathbf{y}_i$ were measured. We assume a mixed linear model of the form

$$y_{ij} = \mu + \alpha_i + S(t_{ij}) + \epsilon_{ij}$$

where μ denotes an overall mean, α_i denotes a random subject effect from a $N(0, \sigma_\alpha^2)$ distribution, $S(t_{ij})$ is a cubic spline function, and the ϵ_{ij} are independent errors from a $N(0, \sigma_e^2)$ distribution. We further assume that the subject effects and the error terms are mutually independent. The above model implies that the correlation between repeated measurements of hormones on the same subject is equal to $\sigma_s^2/(\sigma_s^2 + \sigma_e^2)$.

We chose to use cubic splines to model estradiol levels versus day of cycle because this method provides great flexibility in fitting models, is visually smoth, and requires fewer constants to fit than higher degree splines. Rosenberg et al utilized two parabolic and one

linear function to describe the change in estradiol over the menstrual cycle, with only a single continuity restriction. Thus, although their overall function was continuous, it was not smooth at the two join points.

When fitting a cubic spline model, more join points or knots are better if the variable changes quickly over the covariate space. However, too many knots can lead to over-fitting of the data and more parameters to estimate. Stone (1986) suggested that 5 knots should provide enough flexibility for a reasonable number of degrees of freedom.

Given that the average length of a menstrual cycle is 28 days, we positioned 5 knots at the 5 day intervals: 5, 10, 15, 20, and 25 days prior to next menses. Using the + notation of Smith (1979), let

$$u_+ = u \qquad \text{if} \qquad u > 0$$

$$u_+ = 0 \qquad \text{if} \qquad u \leq 0.$$

Then the cubic spline can be specified as:

$$\begin{aligned} S(t) &= \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \beta_4 (t-5)_+^3 + \beta_5 (t-10)_+^3 \\ &\quad + \beta_6 (t-15)_+^3 + \beta_7 (t-20)_+^3 + \beta_8 (t-25)_+^3. \end{aligned}$$

It follows that the overall mixed linear model has the following form:

$$\begin{aligned} y_{ij} &= \mu + \alpha_i + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_3 t_{ij}^3 + \beta_4 (t_{ij}-5)_+^3 + \beta_5 (t_{ij}-10)_+^3 \\ &\quad + \beta_6 (t_{ij}-15)_+^3 + \beta_7 (t_{ij}-20)_+^3 + \beta_8 (t_{ij}-25)_+^3 + \epsilon_{ij} \qquad (1). \end{aligned}$$

This model assumes that the shape of the function describing the change in estradiol over the menstrual cycle is the same for all subjects, but that subjects can differ with regard to their baseline level on day 0.

Several techniques can be utilized to obtain estimates for the variance components and regression parameters in a mixed linear model, including traditional analysis of variance (ANOVA) methods, maximum likelihood (ML) methods, and restricted maximum likelihood

(REML) methods. REML estimates of the variance components are generally preferred, since ML estimates do not take into account the degrees of freedom used to estimate the fixed effects, which can result in estimates of variance components which are downwardly biased (Laird and Ware, 1982). The ANOVA methodology also has limitations, especially with unbalanced data, such as negative variance estimates, and lack of distributional properties (Searle et al). For these reasons, we used the REML method from the SAS PROC MIXED procedure to fit model (1).

Once the parameters in model (1) are estimated, estradiol levels adjusted for day of cycle can be computed using several approaches. One approach is to calculate the deviation of the subject's observed value from the expected value for that day of the cycle based on the fitted curve:

$$x_{ij} = y_{ij} - \hat{S}(t_{ij}). \tag{2}$$

Similarly, when repeated hormone measurements are available on all subjects and the average hormonal level is used as the exposure, the average adjusted for day of cycle can be calculated as:

$$\bar{x}_{i.} = \{\sum_j y_{ij} - \hat{S}(t_{ij})\}/n. \tag{3}$$

An alternative approach is to use the estimate of the random subject effect, $\hat{\alpha}_i$, from (1) for each $i = 1, ..., N$. The best linear unbiased predictor (BLUP) of α_i is $E(\alpha_i|\mathbf{y}_i, \hat{\beta}, \hat{\sigma}_s^2, \hat{\sigma}_e^2)$, the expected value of α_i, conditional on $\mathbf{y}_i, \hat{\beta}, \hat{\sigma}_s^2$, and $\hat{\sigma}_e^2$, which is also the empirical Bayes estimator of α_i. It can be shown that

$$E(\alpha_i|\mathbf{y}_i, \hat{\beta}, \hat{\sigma}_s^2, \hat{\sigma}_e^2) = \hat{R}_{n_i}(\bar{y}_i - \hat{S}(t_{ij})/n_i), \tag{4}$$

where

$$\hat{R}_{n_i} = \frac{\hat{\sigma}_s^2}{\hat{\sigma}_s^2 + \frac{\hat{\sigma}_e^2}{n_i}}.$$

Note that R_{n_i} can also be interpreted as the reliability coefficient of $\bar{y}_i$.

Estradiol levels adjusted for day of cycle can be computed using one of the above approaches for all subjects and then used as the exposure in the usual logistic (or conditional logistic for matched studies) regression model to evaluate the association between estradiol level and risk of breast cancer among pre-menopausal women. The estimates of the standard errors for the regression parameter will be underestimated, however, since the uncertainty associated with the estimates for the calibration curve are not taken into account. In this paper, we utilize a bootstrap algorithm for obtaining estimates of the confidence intervals which include the variability contributed by estimation of the calibration curve. The algorithm is described in the next section.

3 Results

The primary aim of the NYU Women's Health Study is to determine whether endogenous hormones such as estradiol, are associated with risk of breast cancer. Between March 1985 and June 1991, a cohort of healthy women aged 34-65 years were enrolled at the Guttman Breast Diagnostic Institute, New York. At the time of enrollment and at annual screening visits thereafter, women were asked to donate blood and complete a self-administered questionnaire. Serum samples were frozen and stored for future biological assays. Subsequent cases of breast cancer were identified primarily through active follow-up and confirmed by reviewing medical and pathological records.

In order to limit the costs associated with measuring hormone levels in the cohort, a nested case-control study design was used. For each incident case of breast cancer, individually matched controls were selected at random from the risk set consisting of all cohort members alive and free of breast cancer at the time of diagnosis of the case, and who matched the case on menopausal status at entry, age at entry, and number and approximate dates of blood donations up to the date of diagnosis in the case. For additional details of the study design, see Toniolo et al (1991). The association between endogenous estrogens and breast

cancer in post-menopausal women was reported in Toniolo et al (1995). In this paper, we focus on the association between total estradiol and risk of breast cancer in pre-menopausal women.

A total of 498 estradiol measurements from 367 pre-menopausal control subjects were utilized to fit the calibration curve: 278 subjects had 1 measurement, 60 had 2, 28 had 3, and 4 had 4 measurements. Only measurements obtained less than 35 days prior to next menses were included. Total estradiol levels were log transformed prior to model fitting to adjust for deviations from normality.

The estimated mean curve describing the change in log estradiol level over the menstrual cycle is shown in Figure 1. The parameters in model (1) were estimated as follows: $\hat{\mu} = 4.16, \hat{\beta}_1 = .36, \hat{\beta}_2 = -.029, \hat{\beta}_3 = -.00067, \hat{\beta}_4 = .00061, \hat{\beta}_5 = -.012, \hat{\beta}_6 = .012, \hat{\beta}_7 = -.0077, \hat{\beta}_8 = .0052$. These estimates were then used to calculate levels of total estradiol adjusted for phase of menstrual cycle using the three approaches described above.

The results from fitting conditional logistic regression models to the adjusted total estradiol levels are shown in Table 1. The estimate of the logistic regression coefficient based on the adjusted first and average measurements calculated from (2) and (3), respectively, are similar to the estimate using the unadjusted first measurement because in the original study design, cases were matched to controls according to the phase of menstrual cycle of the first measurement.

In contrast, when the emprical Bayes estimator from (4) was utilized as the exposure, the regression coefficient estimate increased substantially. This increase is not surprising, given that the estimator in (4) can be viewed as an estimator of (3) that has been corrected for measurement error. Whittemore (1989), and Armstrong, Whittemore and Howe (1989), have proposed analogous forms of (4) as a method for correcting for measurement error in linear and logistic regression models. The method, commonly referred to as "Stein shrinkage", involves multiplying an exposure variable measured with error by the reliability of the exposure prior to fitting the regression model to obtain corrected coefficient estimates. In

the absence of confounders, measurement error in the exposure variable will result in relative risk estimates that are attenuated compared to the true relative risk. It follows that methods which correct for measurement error should yield higher estimates of relative risk than the uncorrected estimates, and that the regression coefficent based on the empirical Bayes estimator should be higher than the uncorrected average. It should be noted, however, that this estimator does not completely correct for measurement error because the reliability coefficient in (5) should be adjusted for the matching strata (ref. Kim et al, 1995). That is, the between-subject variance component may be overestimated if it is not computed controlling for the variation due to matching.

The width of the confidence intervals in Table (1) are underestimated because the extra variability due to estimation of the calibration curve is not taken into account. To obtain standard errors which incorporate this additional source of variation, we propose the following bootstrap procedure:

1. Generate a bootstrap sample from the control subjects.

2. Fit model (1) to the bootstrap sample to estimate the parmeters of the calibration curve

3. Generate a bootstrap sample from the matched cases and controls, using the matching stratum as the sampling unit.

4. Adjust the total estradiol measurements for day of menstrual cycle using the estimates from step (2).

5. Fit conditional logisitc regression models to the adjusted total estradiol measurements.

6. Repeat (1)-(4) 1,000 times, which is the approximate minimum number of bootstraps necessary to compute bias-corrected confidence limits.

The simple $(1-\alpha)\%$ confidence interval can be constructed using the $\alpha/2$ and $(1-\alpha/2)$ percentiles of the bootstrap distribution. Bias- corrected confidence intervals should be used when the bootstrap distribution of the regression parameter is asymmetric and when the sample size is small.

4 Conclusions

Our proposed method to adjust for the systematic variability of hormones, such as estradiol, over the menstrual cycle in studies of pre-menopausal breast cancer is an improvement over existing methods because all the repeated measurements for each subject are now utilized in estimating the calibration curve. In addition, the incorporation of cubic splines to describe the cyclical pattern of hormone levels results in a much smoother curve. Finally, this technique is easy to implement and also partially corrects for the problem of errors-in-measurement in the hormone levels.

References

Amstrong, B., Whittemore, A., and Howe, G. Analysis of case-control data with covariate measurement error: application to diet and colon cancer. Statistics in Medicine (1989) 8:1151-1163.

Koenig, K.L., Toniolo, P.G., Bonfrer, P.F., et al. Reliability of serum prolactin measurements. Cancer Epidemiology, Biomarkers and Prevention (1993) 137:1068-1080.

Rosenberg, C.R., Pasternack, B.S., Shore, R.E., et al. Premenopausal estradiol levels and the risk of breast cancer: a new method of controlling for day of the menstrual cycle. Am J Epidemiol (1994) 140: 518-525.

Smith, P.L. Splines as a useful and convenient tool. The American Statistician (1979) 33: 57-62.

Stone, C.J. Comment on Hastie and Tibshirani. Statistical Science (1986) 1: 312-34.

Takatani, O., Okumoto, T., Kosano, H. Genesis of breast cancer in Japanese: A possible relationship between sex hormone binding globulin (SHBG) and serum lipid components. Breast Cancer Res Treat (1991) 18:s27-s29.

Toniolo, P., Pasternack, B., Shore, R., Sonnenschein, e., Koenig, K., Rosenberg, C., Strax, P., Strax, S. Endogendous hormones and breast cancer: A prospective cohort study. Breast Cancer Res Treat (1991) 18:S23-S26.

Toniolo, P., Levitz, M., Zeleniuch-Jacquotte, A., Banerjee, S., Koenig, K., Shore, R., Strax,

P., Pasternack, B. A prospective study of endogenous estrogens and breast cancer in post-menopausal women. Journal of the National Cancer Institute (1995) 87-190-197.

Wu, C.H. Free and protein-bound plasma estradiol-17B during the menstrual cycle. J Clin Endocrinol Metab (1976) 43:436-445.

Whittemore, A.S. Errors-in-variables regression using Stein estimates. The American Statistician (1989); 43:226-228.

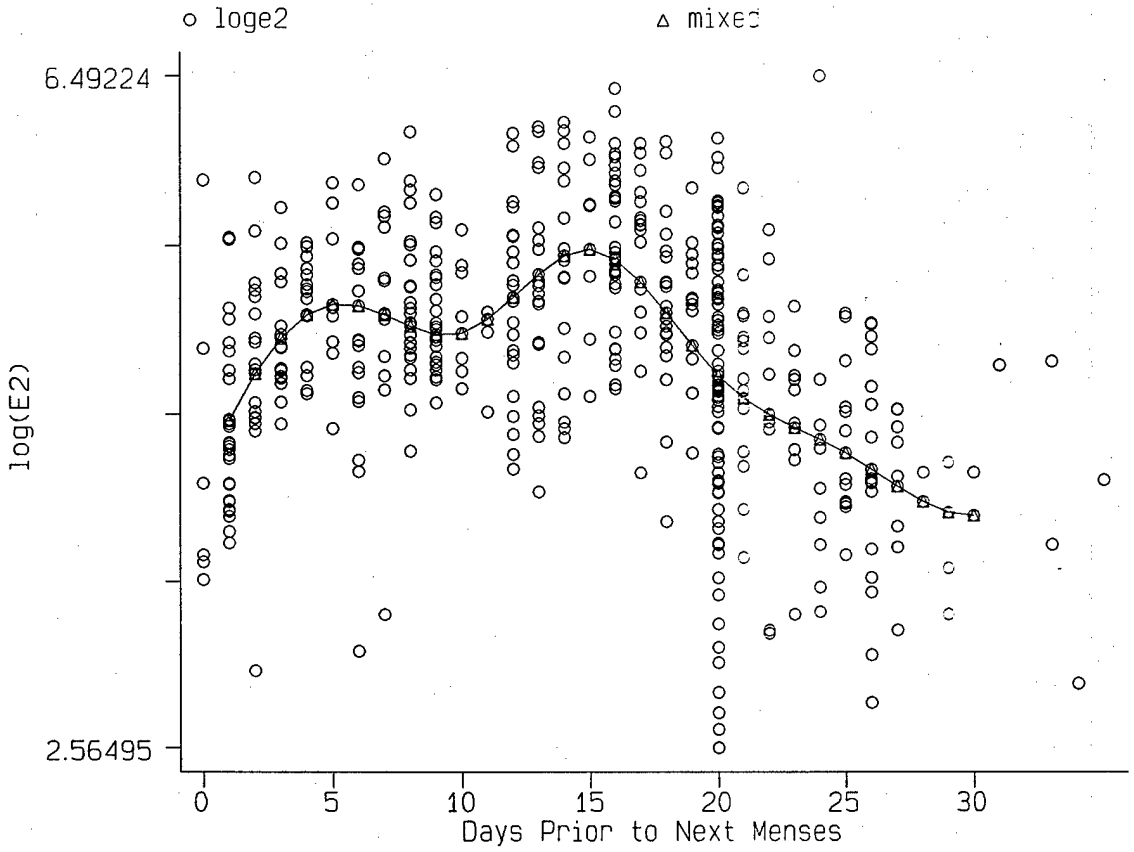

loge2
mixed
6.49224
2.56495
log(E2)
0
5
10
15
20
25
30
Days Prior to Next Menses

Table 1: Logistic Regression Parameter Estimates and 95% Confidence Intervals for the Associations of Total Estradiol Level and Risk of Breast Cancer in Pre-menopausal Women

Exposure Variable	Regression Coefficient	95%C.I.
Total Estradiol		
First measurement	0.19	(-0.23 - 0.61)
Adjusted first measurement	0.26	(-0.17 - 0.70)
Adjusted average	0.17	(-0.28 - 0.63)
Empirical Bayes adjusted average	1.52	(-0.83 - 3.87)

Bibliography

1. Kim, M.Y., and Dubin, N. (1996) Study design and sample size considerations for half-life studies. *Archives of Environmental Contamination and Toxicology* 30:423-429.

2. Kim, M.Y. and Zeleniuch-Jacquotte, A. (1997) Correcting for measurement error in the analysis of case-control data with repeated measurements of exposure. *American Journal of Epidemiology* 145:1003-1010.

3. Zeleniuch-Jacquotte, A., Bruning, P.F., Bonfrer, J.M.G., Koenig, K.L., Shore, R.E., Kim, M.Y., Pasternack, B.S., and Toniolo, P. (1997) Relation of serum levels of testosterone and dehydroepiandrosterone sulfate with risk of breast cancer in postmenopausal women. *American Journal of Epidemiology* 145:1030-1038

4. Lee, E.W. and Kim, M.Y. (1998) The analysis of correlated panel data using a continuous time Markov model. *Biometrics* 54:1638 – 1644.

5. Kim, M.Y. and Shore, R.E. A note on the application of equivalence tests in validation studies of dietary questionnaries. (submitted)

6. Kim, M.Y. Adjusting for cyclical patterns in hormone levels in case-control studies of pre-menopausal breast cancer. (in preparation)